CONVERSION-AWARE FORECASTING OF ALZHEIMER'S DISEASE VIA FEATUREWISE ATTENTION

by

Elvan Karasu

B.S., Chemical Engineering, Boğaziçi University, 2021

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2025

ACKNOWLEDGEMENTS

I would like to express my gratitude to my thesis advisor, Assist. Prof. İnci Meliha Baytaş, for her guidance and support throughout this journey. I have learned so much from her, and I deeply appreciate her time, effort, and belief in me.

I extend my sincere thanks to my jury members Prof. Arzucan Özgür and Prof. Mine Elif Karslıgil, for their insightful feedback.

I gratefully acknowledge the support from TÜBİTAK ARDEB 3501 under grant number 121E594. I also thank Assist. Prof. Doğan Ulus for providing access to the computational resource center, which greatly facilitated my work.

Special thanks to my friend and colleague, Can Koban, for his support and companionship throughout this journey. We faced many challenges together and celebrated our achievements. I am also grateful to all my friends who supported me along the way and made this experience more enjoyable.

I express my deepest thanks to my family, especially my mother, whose boundless love and belief in me have been my greatest source of strength. I dedicate this thesis to my parents, with heartfelt gratitude and love.

The TADPOLE data set constructed by EuroPOND 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. 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 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).

ABSTRACT

CONVERSION-AWARE FORECASTING OF ALZHEIMER'S DISEASE VIA FEATUREWISE ATTENTION

Alzheimer's disease (AD) is a neurodegenerative disorder that causes cerebral atrophy, affecting memory and cognitive functions. An earlier stage, where the symptoms do not overwhelm the patients' daily activities, is known as mild cognitive impairment (MCI). When the underlying cause is AD, MCI patients convert to AD in future stages of the disease. Early detection of the conversion is a vital step in preventative treatment planning. However, identifying this transition is challenging due to the rarity of such events in public datasets. This thesis introduces a deep learning framework to improve conversion detection performance. For this purpose, the proposed architecture is intended to provide insights into the impact of the input biomarkers on conversion detection. It comprises neural blocks to encode attributes and time into a shared space where cross-attention between time and attributes is computed to aggregate the embeddings. While capturing feature importances learned by the model, the temporal information is incorporated into the network with time embeddings. The entire architecture is trained end-to-end to forecast the risk of developing AD. Experiments with two publicly available datasets show promising performance in the early detection of MCI to AD conversion compared to traditional machine learning models and similar deep models. The proposed framework enhances early detection of MCI to AD conversion by effectively integrating temporal dynamics and feature importance. This approach holds promise for improving sensitivity in prediction and supporting more targeted preventative treatment strategies for at-risk individuals.

ÖZET

ÖZNİTELİK DÜZEYİNDE DİKKATTEN YARARLANARAK ALZHEIMER HASTALIĞININ DÖNÜŞÜME DUYARLI TAHMİNİ

Alzheimer hastalığı (AH), hafıza ve bilişsel işlevleri etkileyen ve beyin atrofisine neden olan bir nörodejeneratif bozukluktur. Semptomların hastaların günlük aktivitelerini henüz etkilemediği erken bir aşama, hafif bilişsel bozukluk (HBB) olarak bilinir. Altta yatan neden AH olduğunda, HBB hastaları hastalığın ilerleyen aşamalarında AH'ye dönüşür. Bu dönüşümün erken tespiti, önleyici tedavi planlamasında hayati bir adımdır. Ancak, bu geçişi tespit etmek, kamuya açık veri setlerinde bu tür olayların nadirliği nedeniyle zor bir görevdir. Bu tez, dönüşüm tespitinin performansını geliştirmek için derin öğrenme tabanlı bir çerçeve sunmaktadır. Önerilen mimari, giriş biyobelirteçlerinin dönüşüm tespiti üzerindeki etkisine dair içgörüler sağlamayı hedeflemektedir. Bu yapı, öznitelikleri ve zamanı paylaşılan bir alana kodlamak için sinirsel bloklar içerir; burada, gömüleri birleştirmek için zaman ve özellikler arasında çapraz dikkat hesaplanır. Modelin öğrendiği öznitelik önemleri yakalanırken, zamansal bilgiler de zaman gömme vektörleri aracılığıyla ağa entegre edilmiştir. Tüm mimari, AH gelişme riskini tahmin etmek üzere uçtan uca eğitilmiştir. İki kamuya açık veri setiyle yapılan deneyler, önerilen modelin geleneksel makine öğrenimi ve benzer derin öğrenme modelleriyle karşılaştırıldığında, HBB'den AH'ye dönüşümün erken tespitinde umut verici bir performans sergilediğini göstermektedir. Önerilen çerçeve, zamansal dinamikleri ve öznitelik önemini etkili bir şekilde entegre ederek HBB'den AH'ye dönüşümün erken tespitini geliştirir. Bu yaklaşım, tahminlerde duyarlılığı artırma ve risk altındaki bireyler için daha hedefli önleyici tedavi stratejilerini desteklemek için umut vaat etmektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

d	Dimension of attributes
$d^{'}$	Dimension of time embedding
$\mathbf{e_c}$	Embedding for cognitive features
$\mathbf{e_m}$	Embedding for MRI features
f	Function
$\mathbf{h_i}$	Self-attention output embedding of the ith time step
$\mathbf{k}$	Key vector
$\mathbf{k_c}$	Key vector for cognitive features
$\mathbf{k_m}$	Key vector for MRI features
$\mathbf{q_c}$	Query vector for cognitive features
$\mathbf{q_m}$	Query vector for MRI features
$\mathbf{v}$	Value vector
$\mathbf{v_c}$	Value vector for cognitive features
$\mathbf{v_m}$	Value vector for MRI features
$\mathbf{w}_{K1}$	Projection layer
$\mathbf{w}_{V1}$	Projection layer
$\mathbf{W^K}$	Weight matrix to obtain key vectors
$\mathbf{W}_{K2}$	Projection layer
$\mathbf{W}_{W2}$	Projection layer
$\mathbf{W^Q}$	Weight matrix to obtain query vectors
$\mathbf{W^V}$	Weight matrix to value vectors
$\mathbf{x}$	Input vector
y	Binary label
$\mathbf{z_c}$	Embedding for cognitive features
$\mathbf{z_m}$	Embedding for MRI features
α_c	Attention weights for cognitive features
α_{ij}	Attention weight of the jth dimension at the ith time step

α_m	Attention weights for MRI features
β_0	Logistic regression intercept term
β_i	Logistic regression weights
τ_i	Continuous time point at the ith time step
ω_j	Trainable parameter
φ_j	Trainable parameter

LIST OF ACRONYMS/ABBREVIATIONS

1D	One dimensional
2D	Two dimensional
3D	Three dimensional
ACE	Addenbrooke's Cognitive Examination
ADAS	Alzheimer's Disease Assessment Scale
AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AUC	Area Under the Curve
CDR	Clinical Dementia Rating
CDRSB	Clinical Dementia Rating Sum of Boxes
CN	Cognitively normal
CNN	Convolutional Neural Network
CSF	Cerebrospinal Fluid
FAQ	Functional Activities Questionnaire
FATE-Net	Feature Attentive Network with Time Embeddings
FC	Fully Connected
GRU	Gated Recurrent Unit
ICU	Intensive Care Unit
LR	Logistic regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MCI	Mild Cognitive Impairment
MMSE	Mini-Mental State Examination
MoCA	Montreal Cognitive Assessment
MRI	Magnetic Resonance Imaging
NACC	National Alzheimer's Coordinating Center
NIA-AA	National Institute on Aging and Alzheimer's Association
UDS	Uniform Data Set

OASIS	Open Access Series of Imaging Studies
PET	Positron Emission Tomography
pMCI	Progressive Mild Cognitive Impairment
RF	Random Forest
RAVLT	Rey Auditory Verbal Learning Test
RNN	Recurrent Neural Network
ReLU	Rectified Linear Unit
sMCI	Stable Mild Cognitive Impairment
SVM	Support Vector Machine
TADPOLE	The Alzheimer's Disease Prediction of Longitudinal Evolution
t2v	Time2Vec
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

The use of deep learning in healthcare analytics is a rapidly growing research area that supports the study of medical diagnosis, disease progression, and patient treatment. The analysis of medical data has enabled new opportunities in medicine, enhancing the accuracy and efficiency of clinical decision-making and treatment planning, while also alleviating the workload of clinicians. The integration of machine learning is changing the way clinicians approach diagnostics, as these models can capture patterns and insights from large medical datasets that are hard for traditional methods to uncover. These advancements are particularly valuable in managing complex and chronic diseases, where early detection and personalized treatment are both essential and highly beneficial for patients [1].

Neurodegenerative diseases progress gradually; and if left untreated, they can substantially decrease the life standards of patients [2]. As these diseases worsen over time, timely interventions are crucial. Early diagnosis can significantly delay the onset of more severe conditions, improving patients' quality of life while also reducing the burden on healthcare systems. Among these diseases, Alzheimer's Disease (AD), known as the prevalent cause of dementia, causes a steady degradation in the cognitive and physical abilities of patients [2]. The monitoring of the early stages, which could be diagnosed as Mild Cognitive Impairment (MCI), is critical to detect the risk of converting to AD. MCI is an early form of AD, where the cognitive performances of patients, such as memory and decision-making, are negatively affected [3]. Progression of MCI leads to AD, and predicting this progression ahead of time enables treatment planning to alleviate the symptoms that complicate the daily activities of patients and their loved ones [4].

The diagnosis of AD is done by examining the patient's neuroimaging data, cognitive test scores, genetic information, and blood samples [5]. These diagnostic methods provide a large amount of data; however, interpreting this data manually can be highly

complex. Machine learning algorithms can efficiently process large amounts of data and uncover hidden patterns within them. These algorithms use diagnostic data to model the conversion from MCI to AD. Prediction of the conversion can be posed as a classification task, with patient data represented as a multivariate time series. Consequently, deep learning models are inevitable candidates for solving this classification problem with a temporal nature. Temporal models capture the dynamic changes in variables over time, making them more effective in tracking disease progression [6].

Recurrent neural network (RNN) and its popular variants, such as long short-term memory (LSTM) [7] and gated recurrent unit (GRU) networks [8], are widely used to capture temporal dependencies in temporal patient data. Similarly, transformer networks successfully employed in natural language processing and vision domains can be used for temporal modeling [9]. 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 intensive care unit (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. In the healthcare domain, interpretability becomes even more important, as understanding the decision mechanism of the models enhances their reliability for consideration during clinical decision-making. Trustworthy models support clinicians in managing treatment and monitoring disease progression effectively throughout the course of the disease.

This thesis addresses these challenges by introducing a new framework for AD prediction. The proposed model not only aims to predict the diagnoses accurately, but also to identify the contributing factors in these decisions. We propose an inher-

ently interpretable feed-forward architecture, named Feature Attentive Network with Time Embeddings (FATE-Net), to improve the early conversion prediction between MCI and AD diagnoses. The FATE-Net addresses the explainability of the model's output regarding the input biomarkers. For this purpose, the FATE-Net learns how to attend to each input biomarker by employing self-attention, where the Time2Vec [10] embedding of the visit time is treated as a query. This thesis considers tabular patient data comprising cognitive test scores and volumetric measurements obtained from brain imaging. The model's compatibility with tabular data makes it flexible, as this data format is widely used in clinical practice. This makes it more applicable to real-world settings. Consequently, FATE-Net is not limited to AD, it can be adapted for use in various diseases where tabular data is available.

The key contributions of the proposed approach are summarized below, highlighting its interpretability, temporal modeling, and adaptability for various applications.

- FATE-Net includes a self-attention module that offers interpretability via its self-attention weights. It inherently provides the contributions of the input biomarkers to the prediction. Thus, we can interpret the underlying factors driving the conversion.
- A time embedding module is employed in order to incorporate time information instead of a recurrent structure. Due to the neurodegenerative nature of AD, there is a correlation between the time of the visit and the disease progression. Therefore, the FATE-Net explicitly learns time embedding, which could play a significant role in conversion prediction from MCI to AD.
- The proposed architecture can be trained with cross-sectional patient data. Therefore, FATE-Net is applicable to a scenario where patients have very few visits. It is flexible enough to handle patients with both single and multiple visits simultaneously.
- FATE-Net's feature attention is not specific to AD. It applies to any healthcare task with tabular input biomarkers and time information. It offers interpretability for these biomarkers, making the model more reliable, which is an important

issue in the healthcare domain. FATE-Net can be applied to a wide range of medical conditions that have a temporal nature and require the interpretation of biomarkers.

The efficacy of the proposed FATE-Net is demonstrated through experiments conducted on two publicly available datasets. Results show a significant improvement in recall for MCI to AD conversion prediction compared to commonly used machine learning and deep learning models in the AD prediction literature. Additionally, a comprehensive case study is presented, illustrating the potential real-world application of FATE-Net in a clinical setting. These findings underscore the practicality and impact of the proposed architecture in advancing predictive modeling for neurodegenerative diseases.

2. RELATED WORK

2.1. Background on Alzheimer's Disease

AD is closely linked to the accumulation of abnormal proteins in the brain, specifically beta-amyloid plaques and tau proteins. These protein buildups cause the death of neurons and synapses, which leads to brain volume shrinkage, known as atrophy [11]. This degeneration primarily affects the hippocampus and cortex, regions associated with memory and cognition. As these proteins disrupt connections between neurons, the resulting changes in the brain lead to symptoms such as memory loss and declines in thinking and reasoning abilities in patients [3].

The diagnosis of AD is conducted by examining neuroimaging data, cognitive test scores, genetic markers, and blood samples from the patient [12]. Machine learning algorithms utilize these data to model the conversion from MCI to AD. In this section, we introduce the primary data types used in the detection of AD. After providing a brief background on AD and the features employed, we group the studies in the literature focusing on early prediction of AD into two categories based on their methodologies: traditional machine learning (ML) methods and deep learning models.

2.1.1. Neuroimaging Data

Neuroimaging data comes from magnetic resonance imaging (MRI) and positron emission tomography (PET) scans of the patients. These imaging techniques provide essential insights into the structural and functional changes in the brain that are associated with AD. MRI scans are three-dimensional, constructed from multiple two-dimensional taken across various planes, forming a final 3D representation of the brain. Changes in the volumes of specific regions, such as the hippocampus and ventricles, are indicators of AD, as they reflect patterns of neurodegeneration and atrophy [13, 14]. MRI scans contain information about the volumes of specific brain regions, which are

valuable for tracking disease progression. These volumes can be extracted using third-party software, such as FreeSurfer [15], which processes MRI data to quantify specific brain regions, or they can be fed as raw data into deep models in an end-to-end fashion.

There are many regions in the brain; however, certain areas are particularly relevant to AD and serve as biomarkers due to the changes they undergo as the disease progresses. The ventricles are fluid-filled cavities that surround brain tissue and tend to enlarge as the brain shrinks in AD [14]. The hippocampus, a region associated with memory, is one of the first areas to shrink as patients progress from MCI to AD [16]. Similarly, the entorhinal cortex, which connects the hippocampus to other parts of the brain, experiences volume loss alongside the hippocampus [13]. Additional regions affected by AD include the fusiform gyrus and the middle temporal gyrus [17,18]. They are involved in visual processing tasks, such as recognition and memory [19, 20]. As these areas undergo atrophy, patients typically experience a decline in the specific skills associated with each of these brain regions [21].

2.1.2. Neuropsychological Data

Neuropsychological data includes scores from cognitive assessments administered to patients, such as the Clinical Dementia Rating (CDR), Alzheimer's Disease Assessment Scale (ADAS), Mini-Mental State Examination (MMSE), Rey Auditory Verbal Learning Test (RAVLT), and Functional Activities Questionnaire (FAQ), which are detailed below. In some cases, a single cognitive test may be sufficient for diagnosis, though multiple assessments may also be applied when further confirmation is needed. Generally, cognitive scores are not used in isolation; they are often combined with features from other modalities. Some studies using machine learning models to predict AD have found that cognitive scores show high feature importance [22, 23].

The Clinical Dementia Rating (CDR) is a mental state assessment in which both the subject and their caregiver answer questions across various categories: memory, orientation, judgment, community affairs, home and hobbies, and personal care. Each

category is scored on a scale from 0 to 3, where 0 indicates no impairment, 0.5 indicates questionable impairment, and 1, 2, and 3 indicate mild, moderate, and severe impairment, respectively. While individual category scores carry specific meanings, the total score called the CDR Sum of Boxes (CDRSB), is calculated by summing these scores. The CDRSB ranges from 0 to 18, with lower values indicating no signs of impairment and higher values indicating later stages of dementia [24–26].

The Alzheimer's Disease Assessment Scale (ADAS) is a tool containing both cognitive and non-cognitive assessments [27]. ADAS-Cog-11 is the cognitive assessment part and it includes 11 tasks to measure the participant's memory, language, and comprehension skills [28]. ADAS-Cog-13 is the ADAS-Cog-11 test with two more tasks added, which are about recalling a word after a delay, digit cancellation or a maze problem task. ADAS-Cog-11 scores range from 0 to 70 while ADAS-Cog-13 scores range from 0 to 85 [29]. In general, patients with AD have a higher score than cognitively normal (CN) patients [30].

The Mini-Mental State Examination (MMSE) is a cognitive assessment tool that is easy to apply and typically takes about 10 minutes to complete [31]. It is one of the most commonly used assessments in hospitals [32] and serves as an international standard for both diagnosing AD and monitoring its progression. It is evaluated on a 30 point scale where higher scores reflect better cognitive abilities.

The Rey Auditory Verbal Learning Test (RAVLT) is a test based on word recall [33]. The participant is presented with a list of words and is required to recall them at different time intervals across multiple trials. This allows for an evaluation of the participant's learning, forgetting, and memory skills based on the number of words they can recall [34].

The Functional Activities Questionnaire (FAQ) assesses a patient's ability to perform daily life activities. It includes 10 categories, each scored on a scale from 0 to 3, resulting in a total score ranging from 0 to 30. Lower scores indicate normal indepen-

dence in performing these activities, while higher scores suggest that the participant is significantly dependent and unable to perform them without assistance [35].

Other cognitive tests include Montreal Cognitive Assessment (MoCA) and Addenbrooke's Cognitive Examination (ACE). The tests applied may change according to the participant's education level. These tests are often used in combination during diagnosis. The tests applied may change according to the participant's education level.

2.2. Studies Using Traditional Machine Learning Methods

Traditional ML methods often rely on effective feature selection. Some studies aim to discover the most clinically plausible input biomarkers facilitating the predictive tasks performed by traditional ML methods. In this regard, Irfan *et al.* proposed a feature selection method to discover vital cognitive features for early AD detection [36]. 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 Positron Emission Tomography (PET) biomarkers based on performances of Random Forest (RF) and gradient boosting classifiers [37].

Bucholc *et al.* proposed a hybrid prognostic framework combining unsupervised and supervised machine learning to identify MCI patients at high risk of converting to AD [4]. Their approach involved training Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and ensemble models using 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 [38]. The proposed latent variables are tested with the LR classifier for predicting conversion from MCI to AD.

To study the conversion from MCI to AD, some studies categorize MCI patients based on whether they eventually convert to AD [39, 40]. MCI patients who convert

to AD in later time steps are classified as progressive MCI (pMCI), while those who do not convert are classified as stable MCI (sMCI). 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 [41].

2.3. Studies Using Deep Learning Methods

The studies presented above commonly train shallow models with various input biomarkers, while some studies train deep models to detect patients at high risk of AD conversion. For instance, Tabarestani *et al.* employed LSTM and GRU for predicting future diagnoses, classifying patients across multiple stages, including cognitively normal (CN), MCI and AD [42]. Their models utilized a diverse range of input features, including cognitive scores, numerical data from MRI and PET scans, genetic information, demographic characteristics, and cerebrospinal fluid (CSF) biomarkers. Given the numerical features across all time steps, they approached the conversion from MCI to AD as a time series forecasting problem with three classes (CN, MCI, and AD). Additionally, they also trained the networks to solve the regression task of predicting future mini-mental state examination (MMSE) scores, taking three time steps as input and outputting the labels for up to the next two time steps.

Another study using numerical MRI features for longitudinal forecasting was conducted by Li *et al.* [39], utilizing data from the Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) dataset. They employed an LSTM autoencoder to forecast outcomes based on longitudinal data, including cognitive scores, baseline hippocampal MRI data, and patient demographics [39]. In a similar study, researchers focused on handling missing values by developing a robust LSTM model with modified loss functions and gradients, which enhanced the model's ability to manage incomplete data [43]. 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. As a solution, Nguyen *et al.* proposed an RNN model with a simpler architecture and fewer parameters compared to LSTM networks, designed to inherently manage missing

values [44]. The authors used temporal patient data comprising cognitive and MRI biomarkers, and emphasized the importance of imputing missing values in a clinically plausible way. Their proposed RNN model was trained to predict the missing values of the next clinical visit [44].

An exciting study by El-Sappagh *et al.* proposed to train an LSTM network in a two-stage framework [45]. The first stage predicts the disease progression, while the second stage is a regression task to predict the time of MCI to AD conversion [45]. 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 [45].

Convolutional neural networks (CNNs) and vision transformer-based architectures are commonly used when MRI and PET images are considered [46, 47]. Kim *et al.* [48], 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 [49]. However, the proposed FATE-Net takes tabular input, e.g., cognitive scores and volumetric MRI variables. Therefore, we do not compare the proposed method with frameworks focusing on medical images that could be expensive to acquire. We aim to incorporate the time factor by learning to embed it and inherently discover the relationship between the outcome and the input biomarkers leveraging self-attention. Thus, the FATE-Net automatically provides feature importance while learning to predict disease progression.

Beyond imaging data, multimodal approaches can also integrate numerical features from different domains. Although these numerical features can be directly fed into

deep networks, it is important to recognize that they still represent multiple modalities. Accounting for the distinct nature of each modality may provide additional insights. Therefore, some studies apply multimodal fusion to numerical features and learn the features of each domain separately. For instance, Lee et al. [23] trained four separate GRUs to extract features from four different modalities: demographics, neuroimaging data, neuropsychological assessments, and cerebrospinal fluid (CSF) biomarkers. They tested their model's predictions over the next four time steps following the baseline visit, finding that performance improved when longitudinal data from multiple modalities were integrated.

In exploring solutions for working with tabular data and ensuring interpretability, Arık et al. present an interesting study that introduces a deep learning framework, TabNet, designed to handle tabular data while providing feature importance for interpretability [50]. TabNet employs an encoder-decoder architecture with a self-supervised learning task, where the encoder learns from partially masked input features, and the decoder reconstructs these features, aiming to predict the masked parts. This design helps the model capture meaningful representations by leveraging a representation learning approach. The encoder incorporates a feature transformer, an attentive transformer, and feature masking, while the decoder consists of feature transformer blocks that reconstruct features based on the encoder's output. Additionally, TabNet employs sequential attention to selectively focus on relevant features at each decision step. This approach allows TabNet to efficiently handle tabular data and make decisions that are interpretable at both local and global levels.

Building on insights from the studies that are mentioned above, our proposed framework, FATE-Net, employs a deep learning approach that incorporates temporal dynamics and attention mechanisms to predict disease progression from MCI to AD. Unlike models that focus heavily on imaging data or that face interpretability challenges, FATE-Net takes tabular input, e.g., cognitive scores and volumetric MRI variables. Therefore, we do not compare the proposed method with frameworks focusing on medical images that could be expensive to acquire and also process. We

aim to incorporate the time factor by learning to embed it and inherently discover the relationship between the outcome and the input biomarkers leveraging self-attention. By leveraging self-attention, FATE-Net provides feature importance at each time step, enhancing interpretability while improving sensitivity in conversion prediction. This approach aligns with prior research emphasizing interpretability and practical application in early AD detection, while uniquely focusing on temporal relationships across structured data.

3. METHOD

3.1. Problem Definition

A dataset D contains time series of visits for each patient i. Each visit t is described by a feature vector $X_{i,t} \in \mathbb{R}^d$, where d features represent cognitive test scores and MRI scan data. Each visit also includes a binary label $y_{i,t} \in \{0, 1\}$, indicating a diagnosis of MCI ($y_{i,t} = 0$) or AD ($y_{i,t} = 1$) for patient i at time t. The objective is to design a framework $f : X \rightarrow y$ that forecasts the risk of patient i being diagnosed with AD in a future visit.

The proposed framework, given in Figure 3.1, 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 following sections present the details of the framework.

3.2. Time Encoder

Time is a critical factor in neurodegenerative diseases, as it directly influences the progression and manifestation of symptoms. Temporal changes in an MCI patient's variables, such as cognitive scores or MRI biomarkers, are essential indicators of the disease progression. For example, a decline in memory or hippocampal volume over time often signals a transition towards AD [16]. 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 [51]. Therefore, we could not expect to have a sufficiently large number of visits to capture a temporal pattern effectively. Consequently, using a time encoder, we directly integrate the time factor into the representation learning. Inspired by Kazemi *et al.*, we adopted the Time2Vec

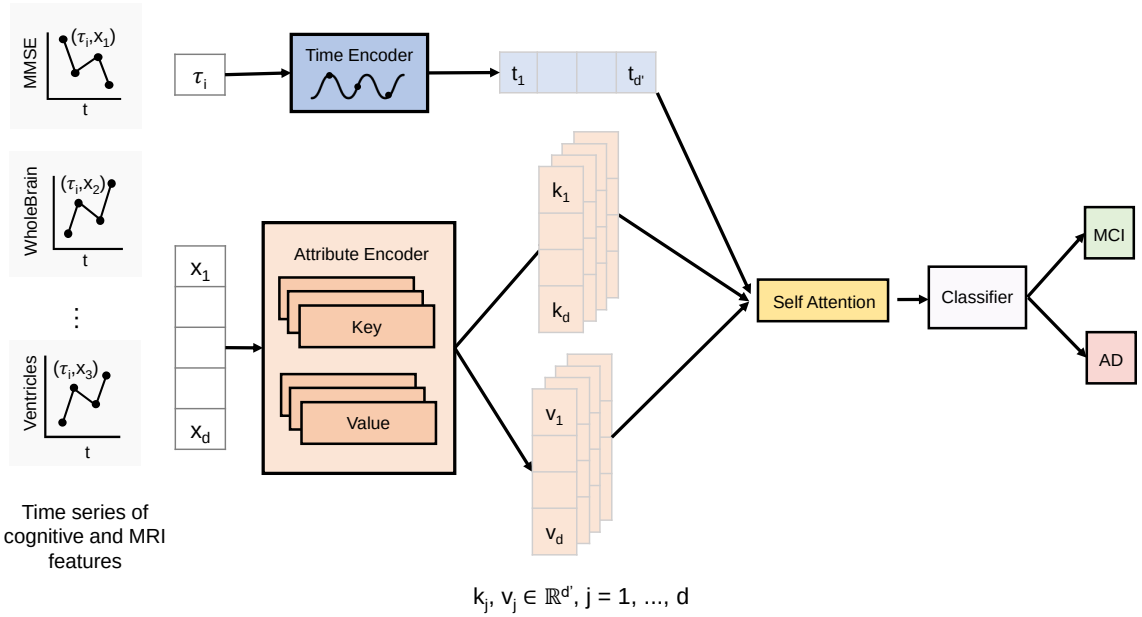

Figure 3.1. FATE-Net architecture. Inspired by Time2Vec [10], the time encoder outputs a time embedding used as the query in the self-attention layer. The attribute encoder maps each input dimension into a space shared by the time embedding. The attribute embeddings are aggregated within the self-attention layer.

representation given below to learn a vector representation from continuous time point τ_i [10]. Time2Vec captures both linear and periodic components of time, as described by

$$\mathbf{t2v}\left(\tau_i\right)[j] = \begin{cases} \omega_j \tau + \varphi_j, & j = 0 \\ \sin\left(\omega_j \tau + \varphi_j\right), & 1 \leq j \leq d^{'} - 1 \end{cases} \tag{3.1}$$

where $\mathbf{t2v}\left(\tau_i\right)[j]$ is the jth element of the $d^{'}$-dimensional time embedding of ith time step, and ω_j and φ_j are trainable parameters. The sinusoidal component, being a periodic function, allows the model to represent periodic behaviors through its frequency ω_j and phase shift φ_j. Meanwhile, the linear component represents the global trends across time.

Time2Vec offers several advantages over traditional methods for representing time. For instance, positional encodings are designed to represent positional order within a sequence; however, they do not account for the time intervals between data points [9]. This limitation can be problematic in medical applications, where sampling intervals are often irregular. In contrast, Time2Vec includes learnable parameters, enabling it to produce embeddings that are more tailored to the specific problem at hand. In neurodegenerative diseases, where changes in variables can be subtle and irregular, Time2Vec provides a robust mechanism to adapt to those variations.

In literature, self-attention with time embeddings has been used to tackle irregularly sampled time series. These methods transform time into vector embeddings to handle with unevenly spaced data observations. For instance, Shukla and Marlin proposed multi-time attention computed over the Time2Vec embeddings of observed time points [52]. They used Time2Vec to transform the time points using learnable weights. After this transformation, they applied self-attention. In their self-attention mechanism, reference time points were treated as queries, whereas observed time points served as keys, and the values corresponded to the observed feature values at each time point. By using the reference time points as queries, the model re-weighted the observed values based on the similarity between the reference time points and the corresponding observed time points.

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 [53]. The proposed FATE-Net does not compute attention weights over the embeddings of all time points. Unlike the previous studies, we treat the learned embedding of the current time point as a query that controls how much attention to allocate to each input biomarker. Section 3.4 presents the proposed self-attention approach.

3.3. 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.

Given an input vector at time step i, $\mathbf{x}_i \in \mathbb{R}^d$, we project each attribute to key and value vectors as

$$\mathbf{k}_{ij} = \text{ReLU}\left(\mathbf{W}_{K2}\text{ReLU}\left(\mathbf{w}_{K1}x_{ij}\right)\right) \tag{3.2}$$

$$\mathbf{v}_{ij} = \text{ReLU}\left(\mathbf{W}_{V2}\text{ReLU}\left(\mathbf{w}_{V1}x_{ij}\right)\right) \tag{3.3}$$

where $\mathbf{k}_{ij}$ and $\mathbf{v}_{ij}$ are the $d^{'}$-dimensional key and value vectors of the jth attribute at ith time step, $\mathbf{w}_{K1} \in \mathbb{R}^{d^{''}}$, $\mathbf{W}_{K2} \in \mathbb{R}^{d^{'} \times d^{''}}$ and $\mathbf{w}_{V1} \in \mathbb{R}^{d^{''}}$, $\mathbf{W}_{V2} \in \mathbb{R}^{d^{'} \times d^{''}}$ are projection layers. Thus, the attribute encoder outputs a key-value pair for each input attribute. The reason behind obtaining attribute-wise key-value pairs is to enable interpreting the predictions in terms of the input biomarker's attention weights.

3.4. Self-Attention Mechanism

The self-attention layer obtains a representation of the patient at the ith 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,

$$\mathbf{h}_i = \sum_{j=1}^{d} \alpha_{ij}\mathbf{v}_{ij} \tag{3.4}$$

yields the final embeddings that can be fed to a classifier. $\mathbf{h}_i \in \mathbb{R}^{d'}$ is the ith time step's final representation before the classifier, and α_{ij} is the attention weight of the jth input dimension at ith time step computed as

$$\alpha_{ij} = \frac{e^{\mathbf{t2v}(\tau_i)^T \mathbf{k}_{ij}/\sqrt{d'}}}{\sum_{k=1}^{d} e^{\mathbf{t2v}(\tau_i)^T \mathbf{k}_{ik}/\sqrt{d'}}} \tag{3.5}$$

3.5. Classification Approaches

Embeddings $\mathbf{h}_i$, representing each visit i, are obtained using the proposed FATE-Net structure, and different methods are explored to utilize these embeddings. In this thesis, the predictive performance of FATE-Net is evaluated using several classifier training options. These classifiers can process either cross-sectional data or temporal data.

The choice of classifier depends on the type of embeddings used. Cross-sectional embeddings capture information from a single visit, such as $\mathbf{h}_i$ for the ith time step. In contrast, temporal embeddings include patterns across multiple visits, such as $\{\mathbf{h}_1, \cdots, \mathbf{h}_T\}$ for T consecutive visits. By applying these approaches, the strengths of various neural network architectures are utilized to effectively predict the conversion from MCI to AD.

3.5.1. Cross-Sectional Embedding-Based Classification

As a simple yet effective baseline, fully connected layers are employed to train a binary classifier using cross-sectional $\mathbf{h}_i$ embeddings, which represent the data from a single visit. This approach does not consider the temporal relationships between visits and instead focuses on the predictive power of static, individual snapshots of data.

The cross-sectional approach establishes a baseline performance for FATE-Net. It evaluates how effectively patients can be classified using a simple neural network and features derived from individual visits.

3.5.2. Temporal Embedding-Based Classification

To capture the progression of biomarkers, cognitive decline from both cognitive test scores and MRI-based features, and temporal relationships within the data, the classification task is extended to use temporal embeddings and train temporal models. Unlike cross-sectional embeddings, this approach processes the sequence of embeddings $\{\mathbf{h}_1, \cdots, \mathbf{h}_T\}$ together, enabling the model to learn temporal patterns that evolve over time. Three different neural network architectures are investigated to process these temporal embeddings:

3.5.2.1. LSTM-Based Classification. LSTM networks are widely used architectures for sequential data, as they process the data recurrently and include gating mechanisms for remembering and forgetting information. These gates enable the model to capture both long-term and short-term dependencies within the sequence. By utilizing the embeddings obtained from FATE-Net, the LSTM cells can effectively process the temporal information to model patterns in the data.

3.5.2.2. 1D-CNN-Based Classification. One-dimensional convolutional neural networks (1D-CNNs) process sequences by detecting local patterns within the embeddings. Unlike LSTMs, which rely on recurrent architectures and gating mechanisms, 1D-CNNs analyze sequences through convolutional operations, making them efficient for capturing short-term trends and identifying anomalies in the features. This capability may enable the model to detect AD conversion at an earlier stage.

3.5.2.3. Transformer-Based Classification. Transformers were originally developed for natural language processing tasks, where their self-attention mechanism enables the model to assign attention weights to different elements in a sequence [9]. In time series data, this mechanism allows the model to focus on specific time points. For this study, it is critical to focus more on MCI visits with a potential for AD conversion rather than visits where AD has already manifested, as capturing the conversion visits is essential.

Transformers can be effective when certain time points carry more relevant information about the patient's condition, allowing the model to attend more to these time points. Unlike LSTMs or 1D-CNNs, transformers process the entire sequence simultaneously rather than sequentially. To preserve the temporal order, positional encodings are added to the embeddings. These encodings enable the model to keep the sequence's order information.

All these networks have different mechanisms for processing sequential data, and their integration with FATE-Net is explored to address the specific problem of AD detection. As temporal models, they respect the sequential structure of the embeddings and can capture dynamic patterns that a static, cross-sectional approach may miss. By using temporal modeling, these networks can extract more information from the temporal context rather than treating each time step independently.

These networks are explored as they are potential options in AD classification [54]. Each of these architectures utilizes the embeddings generated by FATE-Net. The classifiers are trained end-to-end with the time and attribute encoders, using the binary cross-entropy loss function

$$\begin{aligned}\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} &-y_n \log\left(\text{FATE-Net}\left(\mathbf{x}_n, \tau_n\right)\right) \\ &-\left(1-y_n\right)\log\left(1-\text{FATE-Net}\left(\mathbf{x}_n, \tau_n\right)\right)\end{aligned} \tag{3.6}$$

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 $\mathbf{x}_n$ and τ_n are the input feature vector and the time step of the nth visit, respectively.

4. EXPERIMENTS AND RESULTS

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 [51,55], and Uniform Data Set (UDS) version 3.0 from National Alzheimer's Coordinating Center (NACC) [56].

The details about the patients and their visits in the selected cohorts from both datasets are presented in Table 4.1 and Table 4.2, respectively. Patients are categorized into three groups based on their disease progression path. As mentioned in Section 2.2, 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 to AD in their current visit. AD-to-AD visits refer to visits where the patient remains diagnosed with AD.

Table 4.1. Number of patients in TADPOLE and NACC datasets across training, development, and test sets.

TADPOLE

Patient group	Train	Validation	Test	Total
sMCI	330	92	104	526
pMCI	193	47	58	298
AD starter	206	44	67	317
Total	729	183	229	1141

NACC

Patient group	Train	Validation	Test	Total
sMCI	1525	378	463	2366
pMCI	1148	286	376	1810
AD starter	4283	1076	1335	6694
Total	6956	1740	2174	10870

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 CN, MCI, and AD patients with demographics, cognitive test scores, and MRI and PET scans [51, 55]. 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 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 of the patients have fewer than 10 visits.

This thesis focuses only on visits labeled as MCI and AD. Thus, the dataset comprises 1,141 patients with 6,787 visits, which are labeled as MCI and AD, with

Table 4.2. Number of visits in TADPOLE and NACC datasets across training, development, and test sets.

TADPOLE

Visit label	Train	Development	Test	Total
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

Visit label	Train	Development	Test	Total
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

4,308 and 2,479 visits, respectively. Out of these patients, 26% (298 patients) of them start as MCI and convert to AD after some time, which are called progressive MCI (pMCI) patients. Meanwhile, 46% (526 patients) start as MCI and remain MCI, which are called stable MCI (sMCI) patients. The aim is to focus on pMCI patients, especially the visit that they converge to AD, and capture the risk of developing AD.

The patients are split into 5 splits, each split 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 bigger 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 [56]. The National Institute on Aging and Alzheimer's Association (NIA-AA) criterion is used to diagnose AD. The following 14 features are selected in the experiments: NACCAGE, CDRSUM, BILLS, TAXES, SHOPPING, GAMES, STOVE, MEALPREP, EVENTS, PAYATTN, REMDATES, TRAVEL, FAQ, MAXFAQ.

The FAQ, as introduced in Section 2.1.2, is used to assess 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 [57]. 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 5 splits, each split 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 6 visits. As mentioned in Section 4.1.1, missing values are filled with forward and backward filling.

4.2. Preliminary Experiments

This section introduces the methods explored and investigated prior to the development of FATE-Net, focusing on key considerations during its design. The section begins with an exploration of feature importance (Section 4.2.1) and proceeds to examine methods for emphasizing features from various modalities (Section 4.2.2). Each approach is described in detail in the following subsections. In the preliminary experiments, train-dev-test splits distinct from those presented in Section 4.1 are used. The patients in these splits differ from those in the main experiments discussed in Section 4.3 and later, as the splits were independently formed for exploratory purposes.

4.2.1. Exploring Feature Importance

At the initial stage of addressing the AD prediction problem, an exploratory analysis is conducted on the features. For this purpose, a logistic regression model is trained. While logistic regression is a traditional machine learning model, it remains highly effective and practical, particularly due to its ability to provide interpretability through feature weights. As mentioned in Section 4.1.1, we begin with TADPOLE's suggested variables and exclude those with high missing rates. The logistic regression

model is defined as

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}} \tag{4.1}$$

where $P(y = 1 \mid X)$ represents the probability of a positive AD diagnosis given the feature vector X, β_0 is the intercept, β_i are the weights associated with each feature x_i, and n is the total number of features. The weights β_i indicate the contribution of each feature to the final prediction, where positive weights increase the probability of an AD diagnosis, and negative weights decrease it. The magnitude of the weights indicates the relative importance of the features in the prediction.

A logistic regression model is trained on the TADPOLE dataset. Since the goal is to forecast the next visit's label from the current visit, each visit's ground truth label is set as the diagnosis at the next visit. Using this approach, the task is framed as a forecasting problem rather than a standard classification task. To specifically focus on patients converting from MCI to AD, a subset of patients is selected where all individuals begin with an MCI diagnosis and later convert to AD. Among these patients, the visits corresponding to the point of conversion are separated, and classification performance for these points is evaluated and reported as a metric referred to as *conversion recall*. The performance of the logistic regression model, including conversion recall, is presented in Table 4.3.

Looking at Table 4.1, converters constitute 28% of the total patient population. Although this is a small subset, it represents the critical focus of this study, as predicting conversion beforehand would be clinically valuable. While the logistic regression model achieves strong overall metrics, identifying conversion visits remains challenging. While predicting the MCI class is relatively easier for the model, the recall drops by 13% (from 94% to 82%) when we come to the AD class. Conversion recall is even lower (37%), reflecting the model's difficulty in identifying the conversion steps in time, which is the primary clinical concern.

Following the performance evaluation, the weights produced by the logistic regression model are analyzed. The attribute weights for the TADPOLE dataset are

presented in the plot shown in Figure 4.1. For instance, the FAQ variable has a positive weight, which aligns with expectations. A higher FAQ score indicates that the patient struggles with daily activities, increasing the probability of an AD diagnosis [35]. In contrast, most MRI features, such as Hippocampus, Entorhinal, Fusiform, and MidTemp, exhibit negative weights. This is consistent with known pathology, as a decrease in these values corresponds to shrinkage in these brain regions, an indicator of AD, which leads to an increased probability of an AD diagnosis [16, 17].

In terms of magnitude, cognitive features generally show higher feature importance. This raises questions about the role of MRI features, as their weights tend to have lower magnitudes compared to cognitive features. Additionally, some MRI features show unexpected patterns, which might indicate potential interactions or offset effects between features. For instance, larger Ventricles generally increase the probability of an AD diagnosis, while higher WholeBrain volumes decrease it. However, in this case, Ventricles has a negative weight and WholeBrain has a positive weight, which may indicate that these incorrect effects may cancel each other out. Moreover, many MRI features are highly correlated, as these brain regions tend to decrease together as the patient progresses toward AD. This correlation might cause their effects to overlap, resulting in lower weights for each feature.

To evaluate the importance of cognitive and MRI feature groups, we train 2 more models, each excluding one feature group; and examine the effect on performance. The performance of these models, trained using only cognitive features (CDRSB, ADAS11, ADAS13, MMSE, RAVLT_immediate, RAVLT_learning, RAVLT_forgetting, and FAQ) and only MRI features (Ventricles, Hippocampus, WholeBrain, Entorhinal, Fusiform, and MidTemp) is given in Table 4.3.

Looking at Table 4.3, the performance of the models trained using cognitive features and all features is nearly identical. This indicates that cognitive features alone can serve as great predictors for AD classification, as they capture essential aspects of cognitive decline. However, an interesting observation is that MRI features increase

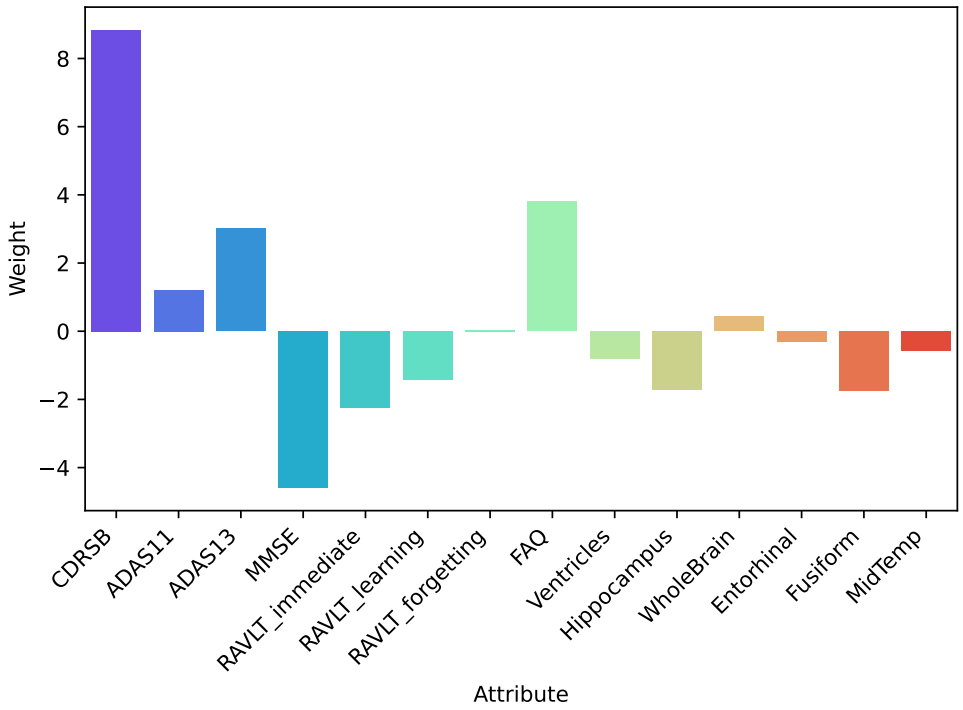

Figure 4.1. Feature weights of the logistic regression model trained on the TADPOLE dataset for predicting the next visit's diagnosis.

Table 4.3. Performance of logistic regression models trained with different feature sets for predicting the next visit's diagnosis. The results represent the average performance from 5-fold cross-validation, with standard deviations in parentheses.

Features	**Precision**		**Recall**		**F1**	**AUC**	**Conversion Recall**
	MCI	**AD**	**MCI**	**AD**			
Cognitive	**0.898 (0.020)**	**0.880 (0.016)**	**0.937 (0.007)**	0.816 (0.025)	**0.892 (0.010)**	**0.877 (0.010)**	0.356 (0.055)
MRI	0.786 (0.029)	0.696 (0.045)	0.842 (0.059)	0.602 (0.047)	0.752 (0.025)	0.722 (0.011)	**0.446 (0.079)**
All	**0.898 (0.020)**	0.879 (0.020)	0.936 (0.011)	**0.817 (0.025)**	0.891 (0.009)	0.876 (0.009)	0.369 (0.083)

conversion recall by 25%, highlighting their impact on detecting individuals likely to progress from MCI to AD. This suggests that MRI features play a critical role in the conversion steps, making them particularly valuable for monitoring disease progression and early intervention strategies.

4.2.2. Fusion of Modalities

4.2.2.1. Traditional Fusion Methods. The impact of MRI features on conversion recall performance highlights the need for further investigation into their role and contribution. We explored these features further to determine whether treating features from different modalities separately could enhance performance. Specifically, we aimed to identify more effective ways to integrate MRI features into temporal models to maximize their contribution. Inspired by the work of Lee et. al, as mentioned in Section 2.3, where separate encoders are used for cognitive and MRI features, we adopted a similar approach. Two distinct LSTM encoders were employed for cognitive and MRI features, referred to as LSTM_c and LSTM_m. These encoders process the cognitive and MRI features separately and generate embeddings for each modality.

Once the embeddings are obtained, various fusion methods can be applied to combine the cognitive and MRI embeddings. Traditional methods include concatenation, where the embeddings are concatenated and passed through a feedforward neural network for classification as MCI or AD. Another approach involves summing the embeddings, either as a simple sum or a weighted sum, where predefined weights are assigned to each modality. These traditional fusion methods, as applied in Lee et al.'s work, were implemented with the LSTMs and a fully connected classifier trained together end-to-end. The performance of these methods on the TADPOLE dataset is presented in Table 4.4.

Looking at the first four rows in Table 4.4, which compare traditional fusion methods, in terms of overall performance, all approaches provide similar F1 and AUC (Area Under the Curve) scores in terms of overall performance. As observed in Table 4.3, where using only MRI features reduced overall performance (F1 or AUC), a similar trend is also seen here: increasing the weight of MRI embeddings leads to a decline in overall performance. Moreover, increasing the weights of MRI embeddings does not improve AD recall or conversion recall in this case. The highest AD recall and conversion recall are achieved using a simple summation. This suggests that the

Table 4.4. Performance of different fusion methods for predicting the next visit's diagnosis. The results represent the average performance from 5-fold cross-validation, with standard deviations provided in parentheses.

Fusion Method	**Precision**		**Recall**		**F1**	**AUC**	**Conversion Recall**
	MCI	**AD**	**MCI**	**AD**			
Concatenation	0.93 (0.01)	0.82 (0.03)	**0.94 (0.01)**	0.78 (0.03)	**0.90 (0.01)**	0.86 (0.02)	0.36 (0.08)
Summation	**0.94 (0.01)**	0.79 (0.04)	0.93 (0.02)	**0.80 (0.04)**	**0.90 (0.02)**	0.87 (0.02)	0.40 (0.09)
Weighted Sum (3:1 Cog:MRI)	**0.94 (0.01)**	0.80 (0.03)	0.93 (0.01)	**0.80 (0.03)**	**0.90 (0.01)**	0.87 (0.02)	0.37 (0.10)
Weighted Sum (1:3 Cog:MRI)	0.92 (0.01)	0.79 (0.04)	0.93 (0.02)	0.76 (0.03)	0.89 (0.02)	0.85 (0.01)	0.36 (0.09)
Self-Attn ($\mathbf{q_c} \cdot \mathbf{k_m}$)	0.93 (0.02)	0.75 (0.07)	0.91 (0.04)	0.78 (0.09)	0.88 (0.02)	0.85 (0.04)	0.45 (0.15)
Self-Attn ($\mathbf{q_m} \cdot \mathbf{k_m}$)	0.93 (0.01)	0.75 (0.05)	0.91 (0.02)	**0.80 (0.03)**	0.89 (0.02)	0.86 (0.02)	**0.47 (0.10)**
Self-Attn ($\mathbf{q_m} \cdot \mathbf{k_m}$) with MRI Decay	0.93 (0.01)	0.76 (0.03)	0.92 (0.02)	0.78 (0.04)	0.89 (0.01)	0.85 (0.02)	0.36 (0.17)
Self-Attn ($\mathbf{q_m} \cdot \mathbf{k_m}$) with FC Embeddings	0.92 (0.01)	**0.81 (0.05)**	**0.94 (0.02)**	0.76 (0.04)	**0.90 (0.03)**	**0.90 (0.01)**	0.38 (0.14)

weights assigned for cognitive and MRI embeddings may not be optimal for this task, raising the question of whether there is a more effective approach for summing these embeddings.

4.2.2.2. Fusion via Self-Attention Mechanism. To address the limitations of traditional summation methods, we explore the self-attention mechanism introduced by Vaswani et al. [9]. By using the self-attention mechanism in this problem, the weights can be learned, determining how much each modality should contribute to the prediction. For this purpose, cognitive and MRI features are first processed by two LSTMs, LSTM_c and LSTM_m, respectively. These LSTMs produce outputs $\mathbf{e}_{c_1} \dots \mathbf{e}_{c_T}$ and $\mathbf{e}_{m_1}, \dots, \mathbf{e}_{m_T}$. These embeddings are generated following the same procedure used for summation and concatenation methods.

After obtaining the embeddings, they are multiplied by trainable weight matrices $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$, resulting in three sets of vectors for each modality: $\mathbf{q}_c$ and $\mathbf{q}_m$ as the query vectors, $\mathbf{k}_c$ and $\mathbf{k}_m$ as the key vectors, and $\mathbf{v}_c$ and $\mathbf{v}_m$ as the value vectors for cognitive and MRI features, respectively. At a given time step t, the dot products between the query vector for cognitive features ($\mathbf{q}_c$) and the key vectors for both cognitive ($\mathbf{k}_c$) and MRI ($\mathbf{k}_m$) features are passed through a softmax function to compute the attention weights $\alpha_{c_1}, \alpha_{m_1}, \ldots, \alpha_{c_i}, \alpha_{m_i}$, as

$$\{\alpha_{c_i}, \alpha_{m_i}\}_{i=1}^{t} = \text{softmax}\left(\{\mathbf{q}_{c_i} \cdot \mathbf{k}_{c_i},\, \mathbf{q}_{c_i} \cdot \mathbf{k}_{m_i}\}_{i=1}^{t}\right) \tag{4.2}$$

where $\mathbf{q}_{c_i}$ and $\mathbf{q}_{m_i}$ are the query vectors for the cognitive and MRI features at the ith time step, respectively. The softmax function generates α_{c_i} and α_{m_i}, which correspond to the cognitive and MRI attention weights at the ith time step.

The attention weights for cognitive features at different time steps, $\alpha_{c_1}, \ldots, \alpha_{c_t}$, are collectively referred to as α_c. These attention weights are multiplied by the value vector of the cognitive features, $\mathbf{v}_c$, to compute the output for the cognitive features, $\mathbf{z}_c$. Similarly, the attention weights for MRI features at different time steps, $\alpha_{m_1}, \ldots, \alpha_{m_t}$, are collectively referred to as α_m. These are multiplied by the value vector of the MRI features, $\mathbf{v}_m$, to compute the output for MRI features, $\mathbf{z}_m$. After multiplying the attention weights with their respective value vectors as

$$\mathbf{z}_c = \alpha_c \cdot \mathbf{v}_c \tag{4.3}$$

$$\mathbf{z}_m = \alpha_m \cdot \mathbf{v}_m \tag{4.4}$$

the outputs for cognitive and MRI features $\mathbf{z}_c$ and $\mathbf{z}_m$ are obtained. There are several ways to use these outputs. One approach is concatenation, which is the one used in this setting. After concatenating $\mathbf{z}_c$ and $\mathbf{z}_m$, the combined representation is fed into a fully connected classifier for MCI and AD classification. Although these features are concatenated in the final step, this method differs from the traditional fusion methods described earlier. Fusion occurs not only through concatenation but also via the cross-attention mechanism applied between $\mathbf{q}_c$ and $\mathbf{k}_m$, introducing nonlinear interactions between the modalities. This combination improves conversion recall but slightly reduces overall performance compared to traditional methods.

This setup is just one of several configurations explored. For example, using the dot product between $\mathbf{q}_c$ and $\mathbf{k}_m$ is a design choice. In another configuration, we replaced this cross-attention with the dot product between $\mathbf{q}_m$ and $\mathbf{k}_m$, applying self-attention within each modality separately. This configuration achieves the highest conversion recall and AD recall, further emphasizing the importance of correctly integrating MRI features into the system. Compared to traditional methods, it significantly improves conversion recall while maintaining overall performance on par with traditional approaches. This indicates that the model performs similarly on the overall level but identifies more conversion cases than traditional methods.

4.2.2.3. Impact of MRI Decay. Following the improved performance observed when the query vector of MRI features is utilized, further modifications to the MRI features are explored. Specifically, the decay of MRI features is considered. Since some MRI features involve imputed values, they may not always reflect the brain's exact conditions at the corresponding time points. This is especially true for MRI, as MRI scans are usually conducted less frequently than cognitive tests. As a result, some MRI volumes in the dataset are filled using values from past visits and MRI scans. We hypothesized that applying a decay to these older features could improve model performance. To test this, a decay factor is introduced to the MRI features by multiplying each feature by $e^{-0.10t}$, where t is the elapsed time since the last MRI scan. For instance, if the features correspond to the original MRI scan that matches the visit, t is 0, and the decay factor becomes 1, having no effect on the original features. However, as time elapses, the decay factor decreases, reducing the contributions of older MRI features.

When MRI decay is applied, as shown in Table 4.4, there is a loss in conversion recall compared to the $\mathbf{q}_m \cdot \mathbf{k}_m$ configuration. This suggests that the decay strategy might not be suitable and highlights the model's reliance on MRI features while predicting conversion cases. While other metrics remain nearly unchanged compared to the $\mathbf{q}_m \cdot \mathbf{k}_m$ case, the conversion recall drops significantly, resulting in its lowest observed value in the table.

Finally, the method for obtaining the embeddings $\mathbf{e}_{c_1} \dots \mathbf{e}_{c_T}$ and $\mathbf{e}_{m_1}, \dots, \mathbf{e}_{m_T}$ used in the self-attention mechanism is modified. Instead of utilizing LSTMs, a fully connected network was employed to generate the embeddings. This approach achieves the highest AUC and F1 scores; however, it results in a decrease in AD recall and conversion recall compared to embeddings obtained using LSTMs.

4.3. Implementation Details

In this section, the implementation details for FATE-Net are given. For forecasting, each visit's ground-truth label is set as the next visit's diagnosis, as it was done for the preliminary experiments. 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 Section 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, for the time encoder, the dimension is initially increased to 32 and then to 64. After applying self-attention using Eqs. 3.3 and 3.5, a 3-layer network with dimensions [64,16,2] is employed for classification. The layers have leaky rectified linear unit (ReLU) with a leakiness of 0.2 and 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. 3.2 and 3.3 for key and value vectors, and the remaining procedure is the same. The results for FATE-Net with and without Time2Vec (t2Vec) embeddings are given in Table 4.5 and Table 4.6 for TADPOLE and NACC datasets. 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 in Table 4.5 and Table 4.6.

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. 3.4. The 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.

4.4. Baselines

The FATE-Net models are compared with models using cross-sectional data, such as SVM, LR, eXtreme Gradient Boosting (XGBoost), and TabNet [50]. The forecasting performances of these models are presented for TADPOLE and NACC datasets in Table 4.5 and Table 4.6, respectively. 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 [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 feedforward 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.

4.5. Forecasting Performance

The performance of cross-sectional and temporal models with and without FATE-Net are presented in Table 4.5 and Table 4.6 for TADPOLE and NACC datasets. 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,

Table 4.5. Performance evaluation of predicting the next visit's diagnosis on the TADPOLE dataset. The results represent the average performance from 5-fold cross-validation, with the standard deviation provided in parentheses.

Model	Precision		Recall		F1	AUC	Conversion
	MCI	AD	MCI	AD			Recall
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)
FATE-Net + LSTM	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)**
FATE-Net + 1D-CNN	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)
FATE-Net + Transformer	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)

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 cross-sectional models and temporal models without FATE-Net, without significantly compromising overall performance. This demonstrates the benefits of using the visit representations from the proposed framework in temporal models.

Table 4.6. Performance evaluation of predicting the next visit's diagnosis on the NACC dataset. The results represent the average performance from 5-fold cross-validation, with the standard deviation provided in parentheses.

Model	Precision		Recall		F1	AUC	Conversion
	MCI	AD	MCI	AD			Recall
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)
FATE-Net + LSTM	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)
FATE-Net + 1D-CNN	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)
FATE-Net + Transformer	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 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 compared to 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.

During the hyperparameter tuning of FATE-Net, different numbers of heads for the self-attention mechanism were tested, and the results are presented in Table 4.7. Although the overall performance metrics, such as F1 and AUC, tend to improve slightly with a higher number of heads, the 1-head model consistently achieves the highest conversion recall for both datasets. Conversion recall is particularly crucial in this study, as it directly impacts the model's ability to identify critical cases of MCI to AD conversion.

Moreover, the 1-head configuration simplifies the interpretability of attention weights, making it easier to analyze the model's decision-making process and draw meaningful insights. By opting for a simpler architecture, the risk of overfitting is also reduced, which is especially important given the moderate size of the datasets used. Although there is a slight trade-off in overall performance compared to models with multiple heads, the marginal decrease in other metrics is acceptable in light of the benefits gained in simplicity and interoperability.

Given these considerations, the 1-head configuration was selected for FATE-Net. This choice prioritizes conversion recall and interpretability to enhance its utility and reliability in clinical settings while avoiding unnecessary increases in computational complexity.

4.6. Attribute Importance

The time and attribute encoders are designed to enhance not only the conversion recall but also the explainability of the model. With the attribute encoder, each attribute at each time step is assigned an attention weight. In this regard, the attention weights are examined for conversion patients to understand the significance of the input biomarkers. 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

Table 4.7. Performance of FATE-Net variations with different numbers of attention heads for predicting the next visit's diagnosis. Results represent the average performance from 5-fold cross-validation, with standard deviations in parentheses.

TADPOLE

Number of heads	Precision		Recall		F1	AUC	Conversion Recall
	MCI	AD	MCI	AD			
1-head	0.86 (0.02)	0.74 (0.03)	0.84 (0.03)	0.76 (0.06)	0.81 (0.02)	0.80 (0.03)	**0.56 (0.08)**
2-heads	**0.88 (0.02)**	**0.80 (0.02)**	0.88 (0.01)	0.78 (0.03)	**0.85 (0.01)**	0.83 (0.01)	0.45 (0.02)
3-heads	**0.88 (0.02)**	**0.80 (0.02)**	**0.89 (0.01)**	**0.79 (0.02)**	**0.85 (0.01)**	**0.84 (0.01)**	0.47 (0.02)
4-heads	**0.88 (0.02)**	**0.80 (0.02)**	0.88 (0.03)	**0.79 (0.04)**	**0.85 (0.01)**	**0.84 (0.01)**	0.49 (0.09)

NACC

Number of heads	Precision		Recall		F1	AUC	Conversion Recall
	MCI	AD	MCI	AD			
1-head	0.77 (0.01)	0.92 (0.01)	0.80 (0.01)	0.90 (0.01)	0.87 (0.01)	0.85 (0.01)	**0.54 (0.02)**
2-heads	**0.79 (0.01)**	**0.93 (0.01)**	**0.84 (0.02)**	**0.91 (0.00)**	**0.89 (0.01)**	**0.88 (0.01)**	0.50 (0.03)
3-heads	0.78 (0.01)	0.92 (0.01)	0.81 (0.02)	**0.91 (0.01)**	0.88 (0.01)	0.86 (0.01)	0.53 (0.02)
4-heads	0.78 (0.01)	**0.93 (0.01)**	0.82 (0.01)	**0.91 (0.01)**	0.88 (0.01)	0.86 (0.00)	0.52 (0.03)

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 Figure 4.2.

All the MRI attributes (Ventricles, Hippocampus, WholeBrain, Entorhinal, Fusiform, and MidTemp) in Figure 4.2 have a 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, which is a cognitive test score. It is worth noting that as a patient progresses from MCI to AD,

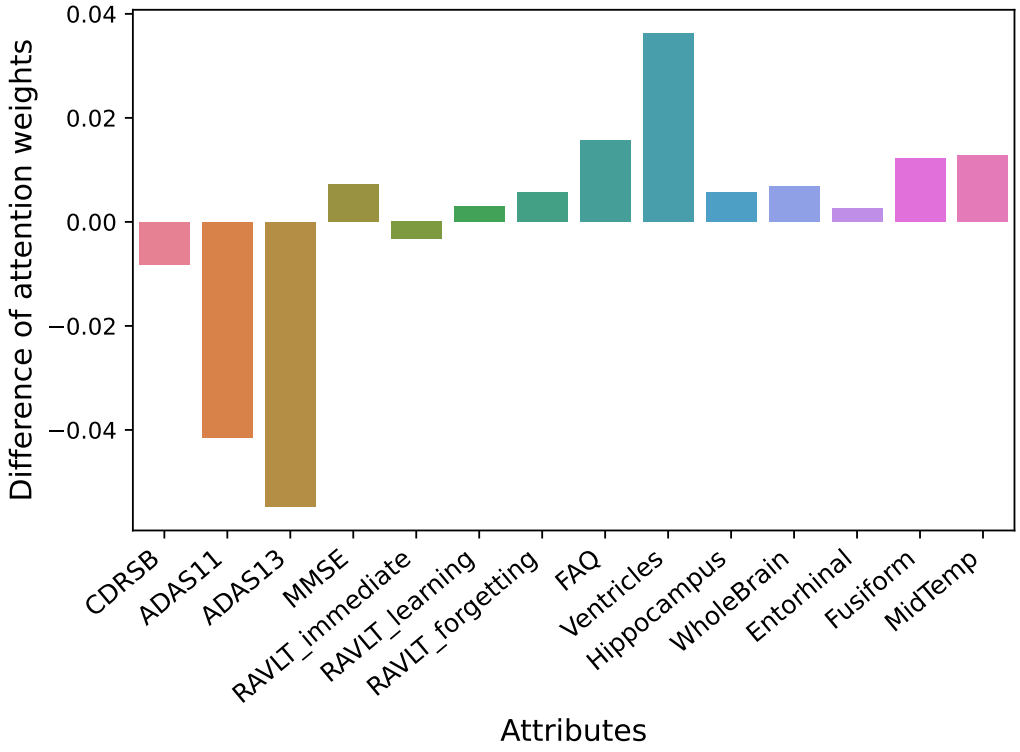

Figure 4.2. Average difference of attention weights between progressive (pMCI) and stable MCI (sMCI) patients.

the ventricles in the brain expand [14]. Therefore, the volume of ventricles serves as a strong indicator of AD, and it is reasonable for the model to pay more attention to this attribute for pMCI patients.

Visits of pMCI and sMCI patients are investigated under two groups: visits where patients remained in the MCI class and where they progressed to AD. The attention weights of the first group are subtracted from that of the second group. In Figure 4.3, the average weight differences are shown. 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 that of Figure 4.2 such that 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

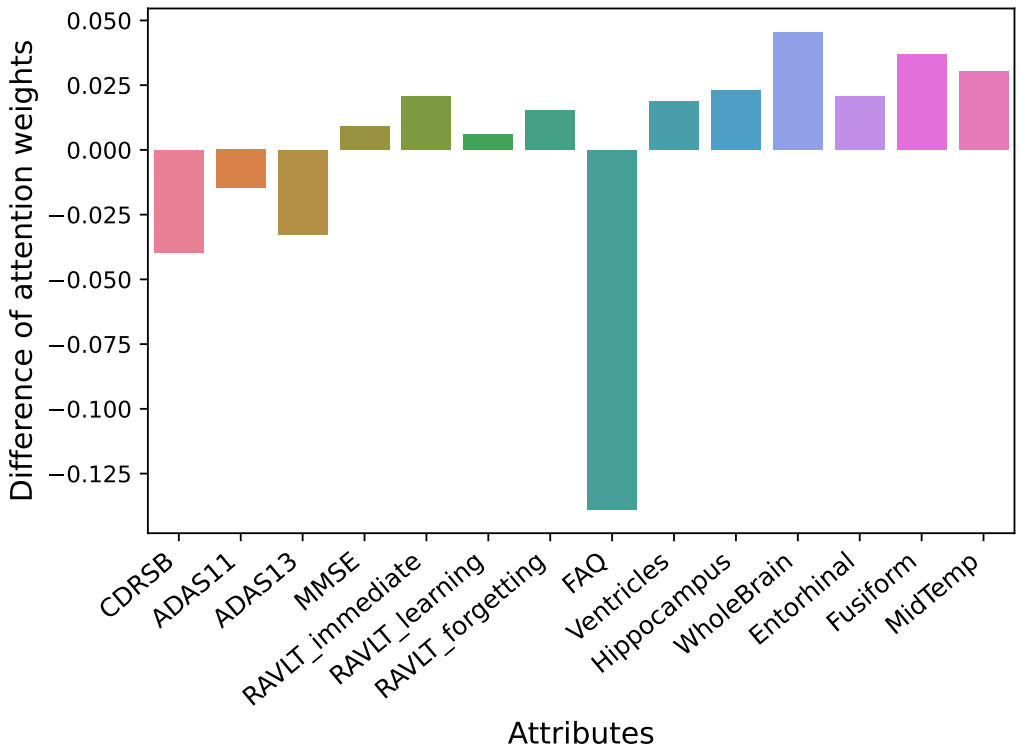

Figure 4.3. Average difference of attention weights between MCI to AD and MCI to MCI visits.

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 cognitive scores [58, 59]. 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.

4.7. 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 4.5. 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 Figure 4.4; where patients on the left are from the TADPOLE dataset, and the ones on the right are from NACC. This analysis provides insights into the models' performance and effectiveness in predicting conversion. The predicted probability of developing AD at the next time step is shown for each current time step. For instance, in TADPOLE, 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. The final two rows show cases where the model's predictions are inaccurate. For instance, the third patient in NACC is predicted to develop AD very early, from the second time step. Although subsequent predictions revert to MCI, at the ninth time step, the model predicts a high probability of conversion to AD at the next step, and the patient indeed converts at the tenth time step.

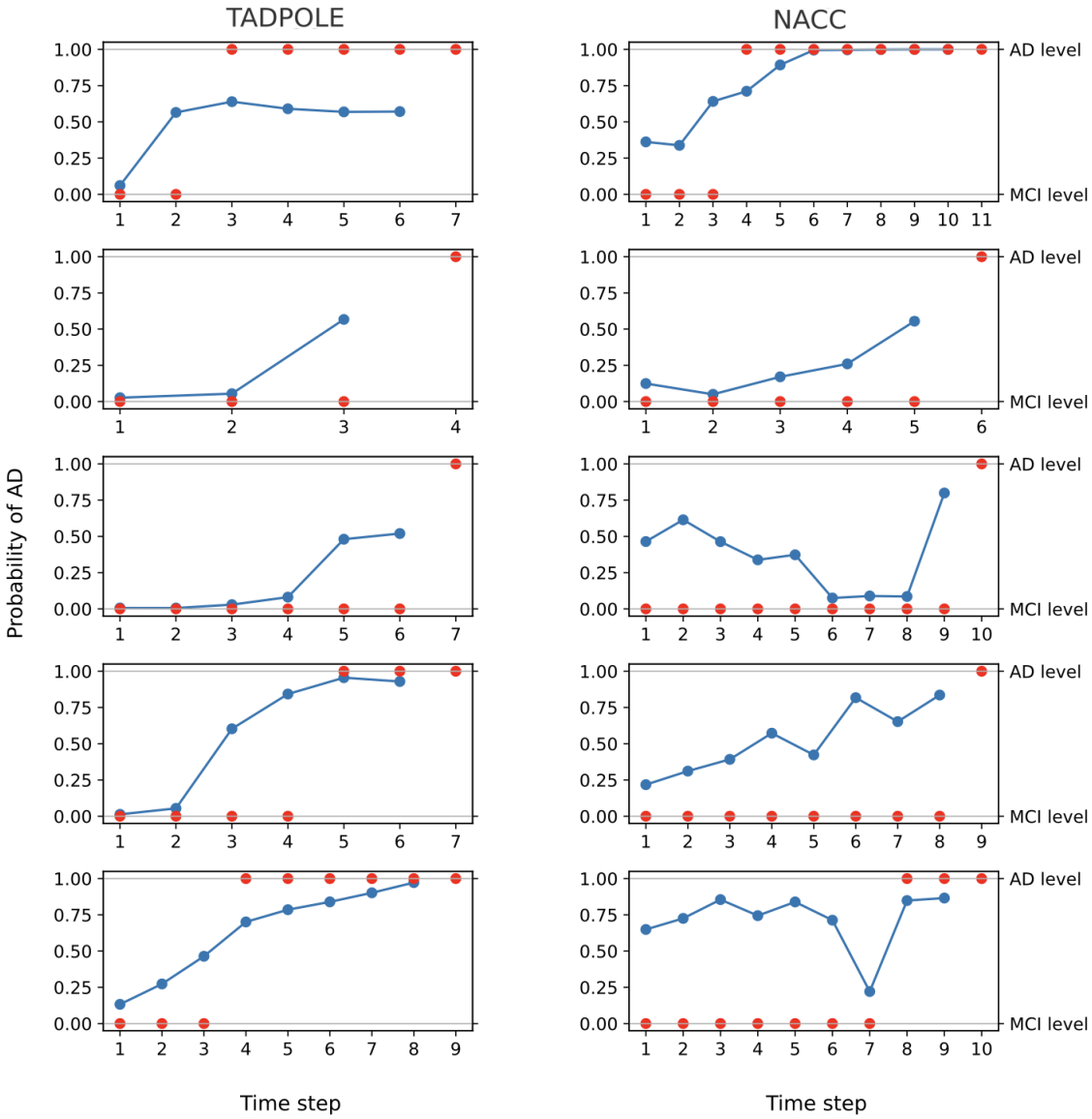

Figure 4.4. Probability of developing AD at the next time step over time for different patient groups using FATE-Net with LSTM on TADPOLE and NACC datasets. The blue lines indicate predicted probabilities for the next time steps, while the red dots represent actual observed MCI and AD labels.

Given the progressive nature of AD, the probability of developing the disease 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 patients in NACC. 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 [60, 61]. Additionally, inconsistencies in testing conditions or measurement errors could also contribute to these variations, making it important to consider multiple assessments over time.

Lastly, even if occurring somewhat early, the model's wrong AD predictions are not necessarily harmful. For instance, the model predicts AD before actual conversion for the fourth patient in TADPOLE; and the third and fourth patients in NACC. For the fourth patient in NACC, the model forecasts conversion at the next time step, and although the conversion does not happen immediately, the patient eventually develops AD. These early predictions can be considered timely alerts for clinicians, providing an early warning. Conversely, false negatives are undesirable. For example, the model underestimates the probability of developing AD for both patients in the fifth row. For the fifth patient in TADPOLE, the probability at the third time step remains below 0.50. For the fifth patient in NACC, the model predicts AD consistently from the beginning; yet incorrectly forecasts MCI in the seventh time step, just before the patient's conversion to AD in the eighth step. Overall, the model demonstrates strong performance, allowing some degree of early false positives for their potential value as early warnings while minimizing false negatives; which remains crucial for accurate forecasting.

5. CONCLUSION

The objective of this thesis is to investigate the conversion from MCI to AD using deep learning methods. The proposed approach, Feature Attentive Network with Time Embeddings (FATE-Net), consists of an attribute encoder and a self-attention mechanism, further enhanced by a time encoder that incorporates visit time into the network using time embeddings. Treating time embeddings as queries in the self-attention mechanism enables identifying the attributes that the model pays attention to at each time 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.

Based on the preliminary experiments and the final framework developed for predicting the risk of AD, several key insights are identified. One significant observation is the critical role of MRI features in predicting conversion. This conclusion is supported by multiple analyses: the performance drop in conversion recall observed in Table 4.3 when MRI features are excluded and a logistic regression model is trained, the sudden decline in conversion recall (with minimal impact on other metrics) when MRI features are modified using a decay factor, as shown in Table 4.4; and the results in Figure 4.3, which highlight the model's greater emphasis on MRI features during conversion steps. The attribute attention of the FATE-Net shows that the MRI biomarkers substantially influence the conversion prediction. Interestingly, this behavior observed in the model aligns with the clinical decision-making process for AD diagnosis. Clinicians usually begin with cognitive tests; and if concerning results are observed, they prioritize MRI scans for further investigation [58, 59]. FATE-Net is able to capture that progressive MCI patients undergo more significant volumetric changes in their brain regions than stable MCI patients [14].

To reach these conclusions, the developed model, FATE-Net, is extensively utilized. The importance and contribution of the self-attention mechanism to this problem become evident during the preliminary experiments with fusion methods, as described in Section 4.2.2. Self-attention not only enhances conversion recall but also maintains overall performance at a level comparable to baseline traditional methods. Additionally, interpretability is a primary consideration, allowing insights into the features the model relies on while making its final decisions. FATE-Net achieves interpretability through its attribute encoder, which maps each feature to a space where attention weights can be analyzed. The time encoder further enhances this by projecting time into a higher-dimensional space, enabling cross-attention between these projections and the attribute features. This mechanism identifies the features most relevant at each time step, contributing to improved performance, especially in terms of conversion recall. These results suggest that the representation of time utilized in FATE-Net can be more effective than a recurrent structure for this task.

Due to its time encoder and attribute encoder modules, FATE-Net provides interpretability at each time step. This feature is especially valuable for identifying MCI patients at risk of AD, as it helps build trust among clinicians by highlighting the features driving the model's predictions. It addresses a key challenge in interdisciplinary applications, where the lack of interpretability in black-box models often limits their practical use. By overcoming this limitation, FATE-Net enhances its practical usability.

FATE-Net is flexible enough to deal with both cross-sectional and temporal data, and it is integrated 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.

Although FATE-Net is originally developed to study MCI to AD conversion, its applicability extends beyond AD. The model is well-suited for problems involving tabular data, particularly when interpretability is essential. It is effective for both

cross-sectional and temporal data, offering flexibility across various applications. Additionally, FATE-Net holds potential for use in other neurodegenerative diseases or entirely different domains where risk prediction or classification tasks require interpretability.

Despite its strengths, FATE-Net has certain limitations. It may not be optimal for datasets with a high number of features. In this thesis, 14 features were used for both datasets, but for high-dimensional data, the model may become inefficient as it projects features into d-dimensional spaces, where d is the number of features. Furthermore, the complexity introduced by the self-attention mechanism increases training time, making the model less efficient than baseline models in this regard.

Future work can address these limitations by adapting FATE-Net for higher-dimensional datasets and reducing its computational complexity. The self-attention mechanism itself offers significant opportunities for further exploration and improvement, as the experiments conducted during fusion methods and with FATE-Net yielded promising results.

Finally, although attention weights are analyzed as average across patients, some individual patient observations reveal fluctuations and unexpected patterns. This suggests there is room for improvement in achieving more reliable interpretability. FATE-Net's interpretability is especially valuable in domains where insights into feature contributions are critical for decision-making. Further enhancements to the cross-attention mechanism can improve its consistency across predictions and provide more reliable results.

REFERENCES

1. Nazir, A., A. Hussain, M. Singh and A. Assad, "Deep Learning in Medicine: Advancing Healthcare with Intelligent Solutions and the Future of Holography Imaging in Early Diagnosis", *Multimedia Tools and Applications*, pp. 1–64, 2024.

2. Association, A., "2019 Alzheimer's Disease Facts and Figures", *Alzheimer's and Dementia*, Vol. 15, No. 3, pp. 321–387, 2019.

3. Association, A., "Mild Cognitive Impairment (MCI) — Symptoms & Treatments", `https://www.alz.org/alzheimers-dementia/what-is-dementia/related_conditions/mild-cognitive-impairment`, accessed on January 23, 2025.

4. Bucholc, M., S. Titarenko, X. Ding, C. Canavan and T. Chen, "A Hybrid Machine Learning Approach for Prediction of Conversion from Mild Cognitive Impairment to Dementia", *Expert Systems with Applications*, Vol. 217, p. 119541, 2023.

5. Clinic, M., "Alzheimer's Disease - Diagnosis and Treatment - Mayo Clinic", `https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/diagnosis-treatment/drc-20350453`, accessed on January 23, 2025.

6. Oxtoby, N. P., *Data-Driven Disease Progression Modeling*, Springer US, New York, 2023.

7. Hochreiter, S. and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

8. Cho, K., B. van Merrienboer, D. Bahdanau and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches", D. Wu, M. Carpuat, X. Carreras and E. M. Vecchi (Editors), *Proceedings of SSST@EMNLP, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–

111, Doha, 2014.

9. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser and I. Polosukhin, "Attention is All You Need", I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Editors), *Advances in Neural Information Processing Systems*, Vol. 30, pp. 6000–6010, Long Beach, California, 2017.

10. Kazemi, S. M., R. Goel, S. Eghbali, J. Ramanan, J. Sahota, S. Thakur, S. Wu, C. Smyth, P. Poupart and M. Brubaker, "Time2Vec: Learning a Vector Representation of Time", *arXiv:1907.05321*, 2019.

11. Selkoe, D. J., "Alzheimer's Disease: Genes, Proteins, and Therapy", *Physiological Reviews*, Vol. 81, No. 2, pp. 741–766, 2001.

12. Jack, C., D. Bennett, K. Blennow, M. Carrillo, B. Dunn, S. Haeberlein, D. Holtzman, W. Jagust, F. Jessen, J. Karlawish, E. Liu, J. Molinuevo, T. Montine, C. Phelps, K. Rankin, C. Rowe, P. Scheltens, E. Siemers, H. Snyder and N. Silverberg, "NIA-AA Research Framework: Toward a Biological Definition of Alzheimer's Disease", *Alzheimer's & Dementia*, Vol. 14, No. 4, pp. 535–562, 2018.

13. McEvoy, L. K. and J. B. Brewer, "Quantitative Structural MRI for Early Detection of Alzheimer's Disease", *Expert Review of Neurotherapeutics*, Vol. 10, No. 11, pp. 1675–1688, 2010.

14. Nestor, S. M., R. Rupsingh, M. Borrie, M. Smith, V. Accomazzi, J. L. Wells, J. Fogarty, R. Bartha and the 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*, Vol. 131, No. 9, pp. 2443–2454, 2008.

15. Fischl, B., "FreeSurfer", *NeuroImage*, Vol. 62, No. 2, pp. 774–781, 2012.

16. Boutet, C., M. Chupin, S. Lehéricy, L. Marrakchi-Kacem, S. Epelbaum, C. Poupon, C. Wiggins, A. Vignaud, D. Hasboun, B. Defontaines, O. Hanon, B. Dubois, M. Sarazin, L. Hertz-Pannier and O. Colliot, "Detection of Volume Loss in Hippocampal Layers in Alzheimer's Disease Using 7 T MRI: A Feasibility Study", *NeuroImage: Clinical*, Vol. 5, pp. 341–348, 2014.

17. Whitwell, J. L., "Progression of Atrophy in Alzheimer's Disease and Related Disorders", *Neurotoxicity Research*, Vol. 18, No. 3, pp. 339–346, 2010.

18. Jobst, K., A. Smith, M. Szatmari, M. Esiri, A. Jaskowski, N. Hindley, B. McDonald and A. Molyneux, "Rapidly Progressing Atrophy of Medial Temporal Lobe in Alzheimer's Disease", *The Lancet*, Vol. 343, No. 8901, pp. 829–830, 1994.

19. Kanwisher, N., J. McDermott and M. M. Chun, "The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception", *Journal of Neuroscience*, Vol. 17, No. 11, pp. 4302–4311, 1997.

20. Pourtois, G., B. de Gelder, A. Bol and M. Crommelinck, "Perception of Facial Expressions and Voices and of Their Combination in the Human Brain", *Cortex*, Vol. 41, No. 1, pp. 49–59, 2005.

21. Eskildsen, S. F., P. Coupé, V. S. Fonov, J. C. Pruessner and D. L. Collins, "Structural Imaging Biomarkers of Alzheimer's Disease: Predicting Disease Progression", *Neurobiology of Aging*, Vol. 36, pp. S23–S31, 2015.

22. Nguyen, H. H., M. B. Blaschko, S. Saarakkala and A. Tiulpin, "Clinically-Inspired Multi-Agent Transformers for Disease Trajectory Forecasting From Multimodal Data", *IEEE Transactions on Medical Imaging*, Vol. 43, No. 1, p. 529–541, 2024.

23. Lee, G., K. Nho, B. Kang, K.-A. Sohn and D. Kim, "Predicting Alzheimer's Disease Progression Using Multi-Modal Deep Learning Approach", *Scientific Reports*, Vol. 9, No. 1, 2019.

24. Berg, L., "Clinical Dementia Rating (CDR)", *Psychopharmacology Bulletin*, Vol. 24, No. 4, p. 637 – 639, 1988.

25. Morris, J. C., "The Clinical Dementia Rating (CDR)", *Neurology*, Vol. 43, No. 11, pp. 2412–2412–a, 1993.

26. Morris, J., C. Ernesto, K. Schafer, M. Coats, S. Leon, M. Sano, L. Thal and P. Woodbury, "Clinical Dementia Rating Training and Reliability in Multicenter Studies: The Alzheimer's Disease Cooperative Study Experience", *Neurology*, Vol. 48, No. 6, p. 1508 – 1510, 1997.

27. Manning, C. A. and J. K. Ducharme, "Handbook of Assessment in Clinical Gerontology", Academic Press, San Diego, 2 edn., 2010.

28. Kueper, J. K., M. Speechley and M. Montero-Odasso, "The Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog): Modifications and Responsiveness in Pre-Dementia Populations. A Narrative Review", *Journal of Alzheimer's Disease*, Vol. 63, No. 2, pp. 423–444, 2018.

29. Mohs, R., D. Knopman, R. Petersen, S. Ferris, C. Ernesto, M. Grundman, M. Sano, L. Bieliauskas, D. Geldmacher, C. Clark and L. Thal, "Development of Cognitive Instruments for Use in Clinical Trials of Antidementia Drugs: Additions to the Alzheimer's Disease Assessment Scale That Broaden Its Scope", *Alzheimer Disease & Associated Disorders*, Vol. 11, No. 2, pp. S13–S21, 1997.

30. Mohs, R. C., D. S. Knopman, R. C. Petersen, S. H. Ferris, C. Ernesto, M. Grundman, M. Sano, L. A. Bieliauskas, D. S. Geldmacher, C. M. Clark and L. J. Thai, "A New Rating Scale for Alzheimer's Disease", *American Journal of Psychiatry*, Vol. 141, No. 11, pp. 1356–1364, 1984.

31. Folstein, M. F., S. E. Folstein and P. R. McHugh, ""Mini-Mental State": A Practical Method for Grading the Cognitive State of Patients for the Clinician", *Journal*

of Psychiatric Research, Vol. 12, No. 3, pp. 189–198, 1975.

32. Gallegos, M., M. L. Morgan, M. Cervigni, P. Martino, J. Murray, M. Calandra, A. Razumovskiy, T. Caycho-Rodríguez and W. L. A. Gallegos, "45 Years of the Mini-Mental State Examination (MMSE): A Perspective from Ibero-America", *Dementia & Neuropsychologia*, Vol. 16, No. 4, pp. 384–387, 2022.

33. Rey, A., *L'examen Clinique en Psychologie*, Presses Universitaries De France, Paris, 1958.

34. Lezak, M. D., *Neuropsychological Assessment*, Oxford University Press, New York, 2004.

35. Pfeffer, R. I., T. T. Kurosaki, J. Harrah, C. H., J. M. Chance and S. Filos, "Measurement of Functional Activities in Older Adults in the Community", *Journal of Gerontology*, Vol. 37, No. 3, pp. 323–329, 1982.

36. Irfan, M., S. Shahrestani and M. Elkhodr, "Early Detection of Alzheimer's Disease Using Cognitive Features: A Voting-Based Ensemble Machine Learning Approach", *IEEE Engineering Management Review*, Vol. 51, No. 1, pp. 16–25, 2023.

37. Franciotti, R., D. Nardini, M. Russo, M. Onofrj and S. L. Sensi, "Comparison of Machine Learning-Based Approaches to Predict the Conversion to Alzheimer's Disease from Mild Cognitive Impairment", *Neuroscience*, Vol. 514, pp. 143–152, 2023.

38. Chang, M. and C. Brainerd, "Predicting Conversion from Mild Cognitive Impairment to Alzheimer's Disease with Multimodal Latent Factors", *Journal of Clinical and Experimental Neuropsychology*, Vol. 44, No. 4, pp. 316–335, 2022.

39. Li, H. and Y. Fan, "Early Prediction of Alzheimer's Disease Dementia Based on Baseline Hippocampal MRI and 1-Year Follow-Up Cognitive Measures Using Deep Recurrent Neural Networks", *IEEE 16th International Symposium on Biomedical*

Imaging, pp. 368–371, Venice, 2019.

40. Lee, G., B. Kang, K. Nho, K.-A. Sohn and D. Kim, "MildInt: Deep Learning-Based Multimodal Longitudinal Data Integration Framework", *Frontiers in Genetics*, Vol. 10, 2019.

41. Kung, T.-H., T.-C. Chao, Y.-R. Xie, M.-C. Pai, Y.-M. Kuo and G. G. C. Lee, "Neuroimage Biomarker Identification of the Conversion of Mild Cognitive Impairment to Alzheimer's Disease", *Frontiers in Neuroscience*, Vol. 15, p. 584641, 2021.

42. Tabarestani, S., M. Aghili, M. Shojaie, C. Freytes, M. Cabrerizo, A. Barreto, N. Rishe, R. E. Curiel, D. Loewenstein, R. Duara and M. Adjouadi, "Longitudinal Prediction Modeling of Alzheimer Disease Using Recurrent Neural Networks", *IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–4, Chicago, Illinois, 2019.

43. Mehdipour Ghazi, M., M. Nielsen, A. Pai, M. J. Cardoso, M. Modat, S. Ourselin and L. Sørensen, "Training Recurrent Neural Networks Robust to Incomplete Data: Application to Alzheimer's Disease Progression Modeling", *Medical Image Analysis*, Vol. 53, pp. 39–46, 2019.

44. Nguyen, M., T. He, L. An, D. C. Alexander, J. Feng and B. T. Yeo, "Predicting Alzheimer's Disease Progression Using Deep Recurrent Neural Networks", *NeuroImage*, Vol. 222, p. 117203, 2020.

45. El-Sappagh, S., H. Saleh, F. Ali, E. Amer and T. Abuhmed, "Two-Stage Deep Learning Model for Alzheimer's Disease Detection and Prediction of the Mild Cognitive Impairment Time", *Neural Computing and Applications*, Vol. 34, No. 17, pp. 14487–14509, 2022.

46. Gao, X., H. Cai and M. Liu, "A Hybrid Multi-Scale Attention Convolution and Aging Transformer Network for Alzheimer's Disease Diagnosis", *IEEE Journal of*

Biomedical and Health Informatics, Vol. 27, No. 7, pp. 3292–3301, 2023.

47. Chen, Q., Q. Fu, H. Bai and Y. Hong, "Longformer: Longitudinal Transformer for Alzheimer's Disease Classification With Structural MRIs", *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3575–3584, Waikoloa, HI, 2024.

48. Kim, J. S., J. W. Han, J. B. Bae, D. G. Moon, J. Shin, J. E. Kong, H. Lee, H. W. Yang, E. Lim, J. Y. Kim, L. Sunwoo, S. J. Cho, D. Lee, I. Kim, S. W. Ha, M. J. Kang, C. H. Suh, W. H. Shim, S. J. Kim and K. W. Kim, "Deep Learning-Based Diagnosis of Alzheimer's Disease Using Brain Magnetic Resonance Images: An Empirical Study", *Scientific Reports*, Vol. 12, No. 1, 2022.

49. Subramanyam Rallabandi, V. and K. Seetharaman, "Deep Learning-Based Classification of Healthy Aging Controls, Mild Cognitive Impairment and Alzheimer's Disease Using Fusion of MRI-PET Imaging", *Biomedical Signal Processing and Control*, Vol. 80, p. 104312, 2023.

50. Arik, S. O. and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning", *AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 6679–6687, Palo Alto, California, 2021.

51. Petersen, R., P. Aisen, L. Beckett, M. Donohue, A. Gamst, D. Harvey, C. Jack, W. Jagust, L. Shaw, A. Toga, J. Trojanowski and M. Weiner, "Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical Characterization", *Neurology*, Vol. 74, No. 3, pp. 201–209, 2010.

52. Shukla, S. N. and B. Marlin, "Multi-Time Attention Networks for Irregularly Sampled Time Series", *International Conference on Learning Representations*, Vienna, 2021.

53. Zhang, X., S. Li, Z. Chen, X. Yan and L. R. Petzold, "Improving Medical Predic-

tions by Irregular Multimodal Electronic Health Records Modeling", *International Conference on Machine Learning*, pp. 41300–41313, PMLR, Honolulu, Hawaii, 2023.

54. Wang, R., L. Zeng, H. Li, L. Lu, J. Li and X. Luo, "Alzheimer's Disease Classification Based on One Dimensional Convolutional Neural Network", *8th International Conference on Digital Home (ICDH)*, pp. 288–293, Los Alamitos, CA, 2020.

55. Marinescu, R., N. Oxtoby, A. Young, E. Bron, A. Toga, M. Weiner, F. Barkhof, N. Fox, P. Golland, S. Klein and D. Alexander, "TADPOLE Challenge: Accurate Alzheimer's Disease Prediction Through Crowdsourced Forecasting of Future Data", *Predictive Intelligence in Medicine: Second International Workshop*, Vol. 11843, pp. 1–10, Springer, Shenzen, 2019.

56. Besser, L., W. Kukull, D. Knopman, H. Chui, D. Galasko, S. Weintraub, G. Jicha, C. Carlsson, J. Burns, J. Quinn, R. Sweet, K. Rascovsky, M. Teylan, D. Beekly, G. Thomas, M. Bollenbeck, S. Monsell, C. Mock, X. Zhou, N. Thomas, E. Robichaud, M. Dean, J. Hubbard, M. Jacka, K. Schwabe-Fry, J. Wu, C. Phelps and J. Morris, "Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set", *Alzheimer's Disease & Associated Disorders*, Vol. 32, No. 4, pp. 351–358, 2018.

57. Vik, A., M. Kociński, I. Rye, A. Lundervold and A. Lundervold, "Functional Activity Level Reported by an Informant is an Early Predictor of Alzheimer's Disease", *BMC Geriatrics*, Vol. 23, No. 1, p. 205, 2023.

58. Johnson, K., N. Fox, R. Sperling and W. Klunk, "Brain Imaging in Alzheimer Disease", *Cold Spring Harbor Perspectives in Medicine*, Vol. 2, No. 4, p. a006213, 2012.

59. Szczesniak, D., J. Rymaszewska, A. Zimny, M. Sasiadek, K. Połtyn-Zaradna, E. Smith, K. Zatońska, T. Zatoński, S. Rangarajan, S. Yusuf and A. Szuba, "Cere-

bral 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*, Vol. 43, No. 1, pp. 279–295, 2020.

60. Blair, M., K. Coleman, S. Jesso, V. Desbeaumes Jodoin, K. Smolewska, E. Warriner, E. Finger and S. H. Pasternak, "Depressive Symptoms Negatively Impact Montreal Cognitive Assessment Performance: A Memory Clinic Experience", *Canadian Journal of Neurological Sciences*, Vol. 43, No. 4, p. 513–517, 2016.

61. Bartfai, A., M. Åsberg, A. Beser, K. Sorjonen, A. Wilczek and S. Warkentin, "Impaired Cognitive Functioning in Stress-Induced Exhaustion Disorder: A New Tablet-Based Assessment", *BMC Psychiatry*, Vol. 21, No. 1, p. 459, 2021.