DEEP LEARNING METHODS FOR PREDICTING ALZHEIMER'S DISEASE PROGRESSION: TRANSFORMER AND GRAPH NEURAL NETWORK APPROACHES

by

MAHDI MOGHADDAMI

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

2026

Oakland University
Rochester, Michigan

Thesis Advisory Committee:

Mohammad-Reza Siadat, Ph.D., Chair
Tianle Ma, Ph.D.
Jia Li, Ph.D.

Acknowledgments

This thesis is based on research conducted at Oakland University. I am grateful to my advisor, Dr. Mohammad-Reza Siadat, and collaborators for their guidance and support throughout this research. Special thanks to Clayton Schubring for his invaluable contributions to the Transformer-based research, and to Austin T. Toma, Connor P. Laming, and Dr. Huirong Fu for their contributions to the GNN-based approach.

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder affecting millions globally. Early detection of disease progression is critical for timely intervention. This thesis presents two complementary deep learning approaches for predicting AD progression using longitudinal data. The first approach employs a Transformer-based model that leverages visit history features, including cognitive test scores and imaging features. The second approach uses a history-aware graph neural network (HA-GNN) that operates on functional connectivity derived from resting-state functional MRI data. Both approaches address key challenges in longitudinal prediction, including irregular visit spacing, missing data, and class imbalance. Results demonstrate that Transformer-based models achieve superior performance on multi-class diagnosis prediction (82.4% accuracy), particularly for identifying disease converters. The HA-GNN model achieves 82.9% accuracy in binary conversion prediction, offering potential for early intervention. These findings emphasize the importance of model selection in predicting cognitive decline, with implications for clinical decision-making and patient outcomes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Alzheimer's disease (AD) is the most prevalent cause of dementia and represents a significant public health challenge worldwide [Scheltens et al., 2016]. Nearly 7 million Americans age 65 and older are living with AD, and it is projected that by 2050, this number may grow to 12.7 million as the aging population increases. AD was the fifth-leading cause of death among individuals aged 65 and older in 2021 in the United States. The disease results in devastating consequences not only for patients but also for their families and society, causing substantial clinical, social, and economic burden.

AD is a progressive neurodegenerative disorder characterized by gradual cognitive decline that unfolds over several years. The disease typically progresses through distinct cognitive stages: from a cognitively normal (CN) state, to mild cognitive impairment (MCI), and finally to advanced dementia (AD). Mild cognitive impairment is a precursor state to AD, with approximately 10-15% of individuals with MCI developing AD every year. Unfortunately, most diagnoses of Alzheimer's disease occur in the moderate to late stages, at which point the optimal time for treatment intervention has often already elapsed.

Since no cure has been found for AD, early detection and accurate prediction of disease progression are critical for timely intervention [The Lancet Editors, 2020]. Recent evidence suggests that disease-modifying treatments can slow cognitive decline if administered early, particularly during the MCI stage. Therefore, developing methods that can accurately predict disease progression—specifically, forecasting the trajectory of cognitive decline and identifying when CN individuals will develop MCI and when MCI individuals will develop AD—has become a vital research objective.

1.1 Motivation and Scope

The primary challenge in AD prediction is developing models that can leverage longitudinal patient data to accurately forecast cognitive transitions. Traditional approaches have focused on cross-sectional classification of cognitive states using static imaging or clinical measures. However, longitudinal prediction offers significant advantages: it captures the individual's disease trajectory and progression rate, providing more clinically relevant information for prognosis. Several factors complicate longitudinal prediction:

1. **Irregular Visit Spacing**: Patients do not visit clinics at regular intervals, creating gaps of varying lengths in the temporal sequence.

2. **Missing Data**: Not all measurements are taken at every visit due to the invasiveness of certain procedures or scheduling issues.

3. **Class Imbalance**: Converter subjects (those transitioning to more severe cognitive states) are less common than stable subjects, creating imbalanced datasets.

4. **High Dimensionality**: Brain imaging data contains high-dimensional information that must be effectively integrated with clinical measures.

5. **Temporal Dynamics**: The model must capture how features evolve over time and how this evolution relates to disease progression.

With recent advances in machine learning and deep learning, researchers have increasingly used computer-aided diagnosis (CAD) systems for AD detection and prediction. Deep learning architectures have shown promise in capturing complex temporal and spatial patterns in clinical and imaging data. Two architectural paradigms have emerged as particularly promising: Transformer-based models for processing structured tabular data and Graph Neural Networks (GNNs) for processing network-structured imaging data.

1.2 Thesis Contributions

This thesis presents two complementary deep learning approaches for predicting Alzheimer's disease progression:

Approach 1: Transformer-Based Longitudinal Model (CHIL 2025) We, Moghaddami et al. [2025] propose a Transformer model that predicts the stage of AD progression at a subject's next clinical visit using features from a sequence of prior visits. The model leverages structured clinical data including cognitive test scores, MRI volumetric measures, PET biomarkers, and cerebrospinal fluid markers. Key contributions include:

- Rigorous comparison of Transformer-based models to recurrent neural networks (LSTM, GRU, minimalRNN)

- Assessment of model performance across different prior visit history lengths

- Analysis of feature importance across different data modalities
- Demonstration of strong performance particularly in identifying converter subjects

Approach 2: History-Aware Graph Neural Network (SPIE 2026) We propose a history-aware graph neural network (HA-GNN) that combines graph convolutional networks with recurrent neural networks to predict disease conversion using functional connectivity patterns from resting-state fMRI. Key contributions include:

- Integration of spatial brain network representation (via GCN) with temporal sequence modeling (via RNN)
- Explicit encoding of irregular visit intervals for longitudinal prediction
- Two-stage training approach with GCN pretraining on single-visit diagnosis classification
- Competitive performance on early-stage detection (CN to MCI transitions)

Both approaches address the challenge of predicting disease progression in the presence of missing data, irregular visit spacing, and class imbalance. They demonstrate how different data modalities (structured clinical features vs. neuroimaging-derived connectivity) and architectural choices (attention-based vs. graph-based) can be effectively leveraged for this critical clinical application.

1.3 Thesis Organization

This thesis is organized as follows:

- **Chapter 2 - Background**: Provides context on Alzheimer's disease pathology, previous work in AD prediction, deep learning architectures, and neuroimaging modalities.

- **Chapter 3 - Transformer-Based Longitudinal Prediction**: Presents the first approach, including data description, preprocessing, model architecture, training procedures, and comprehensive results demonstrating the effectiveness of Transformers for multi-class diagnosis prediction.

- **Chapter 4 - Graph Neural Network Approach**: Presents the second approach using functional connectivity and GNNs, including functional connectivity extraction, model architecture combining GCN with RNNs, and results on binary conversion prediction.

- **Chapter 5 - Comparative Discussion**: Compares and contrasts the two approaches, discussing their respective strengths, weaknesses, complementary nature, and clinical implications.

- **Chapter 6 - Conclusion and Future Work**: Synthesizes findings across both approaches and discusses future research directions.

1.4 Key Findings Overview

The Transformer-based model demonstrates superior performance on multi-class diagnosis prediction with 82.4% accuracy and 0.920 mAUC on the balanced dataset. Notably, the model significantly outperforms RNN baselines in identifying converter sequences, particularly those with shorter visit histories. Statistical analysis confirms these improvements are significant, even compared

to the best-performing RNN baseline (GRU).

The HA-GNN model achieves 82.9% accuracy on overall prediction with particularly impressive performance on CN to MCI conversions (68.8% accuracy), the earliest and most clinically relevant transition. The model demonstrates the ability to capture meaningful patterns from functional connectivity networks and maintain information across irregular visit intervals. These complementary approaches highlight how model architecture selection and feature modality choice significantly impact performance on this challenging clinical task. The thesis demonstrates that both structured clinical features and neuroimaging-derived connectivity patterns contain predictive information about cognitive decline, with different architectures excelling at different aspects of the prediction task.

Chapter 2

Background

This chapter provides the foundational knowledge necessary to understand the two deep learning approaches presented in this thesis. We cover the pathology and diagnosis of Alzheimer's disease, review previous work in AD prediction, introduce the deep learning architectures employed, and discuss the neuroimaging modalities used.

2.1 Alzheimer's Disease

2.1.1 Pathology and Progression

Alzheimer's disease is a progressive neurodegenerative disorder characterized by the accumulation of abnormal protein aggregates in the brain, specifically amyloid-beta plaques and tau tangles. These pathological changes lead to neuronal dysfunction, synaptic loss, and ultimately neuronal death. The disease results in cognitive decline affecting memory, thinking, and behavior, eventually interfering with daily tasks.

The cognitive decline in AD unfolds gradually over years, typically progressing through three stages:

1. **Cognitively Normal (CN)**: Individuals have normal cognitive function with no objective evidence of cognitive impairment, though they may have subjective memory concerns.

2. **Mild Cognitive Impairment (MCI)**: Individuals experience objective cognitive decline that is noticeable to themselves and others, but they retain the ability to function independently in daily activities. MCI is considered a precursor state to AD, with 10-15% of MCI individuals converting to AD annually.

3. **Alzheimer's Disease (AD)**: Individuals experience significant cognitive decline that interferes with daily functioning and independence, typically including memory loss, confusion about time or place, and changes in mood or behavior.

The progression from CN to MCI to AD typically occurs over 7-10 years, though the rate varies considerably among individuals. Some individuals with MCI remain stable for years, while others progress to AD relatively quickly.

2.1.2 Diagnosis and Biomarkers

Diagnosis of AD involves clinical assessment combined with biomarker evaluation [Jack et al., 2008]. The amyloid-beta, tau, and neurodegeneration (ATN) framework provides a standardized approach to defining AD pathology. Key biomarkers include:

- **Cognitive and Functional Tests**: MMSE (Mini-Mental State Examination), MOCA (Montreal Cognitive Assessment), ADAS-Cog (Alzheimer's Disease Assessment Scale), RAVLT (Rey Auditory Verbal Learning Test), CDR (Clinical Dementia Rating), FAQ (Functional Activities Questionnaire).
- **Structural Imaging (MRI)**: Measures of brain atrophy including whole brain volume, hippocampal volume, entorhinal cortical volume, and ventricular expansion.
- **Molecular Imaging (PET)**: Amyloid-beta and tau burden quantified using positron emission tomography.
- **Cerebrospinal Fluid (CSF)**: Biomarkers including amyloid-beta 42, phosphorylated tau, and total tau.
- **Functional Imaging (fMRI)**: Functional connectivity patterns derived from resting-state fMRI showing alterations in brain network organization.

2.2 Previous Work in Alzheimer's Disease Prediction

2.2.1 Traditional Approaches

Earlier approaches to AD prediction primarily focused on cross-sectional classification of cognitive states using structural features. These included:

- Support Vector Machines (SVM) applied to cognitive test scores and imaging measures

- Statistical models predicting specific cognitive outcomes at defined follow-up intervals
- Univariate analyses of individual biomarkers

While these approaches provided useful baseline results, they did not effectively leverage the temporal structure of longitudinal data or capture the dynamic nature of disease progression.

2.2.2 Deep Learning Approaches

Recent work has increasingly applied deep learning to AD prediction, recognizing the ability of these methods to learn complex patterns from high-dimensional data:

- **Recurrent Neural Networks**: LSTM, GRU, and other RNN variants have been applied to longitudinal clinical sequences, showing improvement over traditional methods.
- **Convolutional Neural Networks**: CNN-based approaches have been used for feature extraction from structural and functional MRI images.
- **Hybrid Approaches**: Combinations of CNN for spatial feature extraction and RNN for temporal modeling have shown promise in leveraging both imaging and temporal information.
- **Transformer Models**: Attention-based Transformer architectures have recently been applied to longitudinal health data, showing competitive results compared to RNNs on structured tabular data.

- **Graph Neural Networks**: GNN-based approaches have been applied to functional connectivity networks derived from fMRI, showing the value of explicitly modeling brain network structure.

2.3 Deep Learning Architectures

2.3.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of neural networks designed to process sequential data. They maintain a hidden state vector that carries information forward through the sequence, allowing them to capture temporal dependencies.

The basic RNN update rule at time step t is:

$$h_t = \tanh(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{2.1}$$

where x_t is the input at time t, h_t is the hidden state, and W_{xh}, W_{hh} are weight matrices.

However, basic RNNs suffer from vanishing gradient problems when processing long sequences. This led to the development of more sophisticated architectures:

Long Short-Term Memory (LSTM): LSTMs [Hochreiter and Schmidhuber, 1997] introduce memory cells with explicit mechanisms for learning what

information to store, retain, and forget:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.3}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2.4}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.6}$$

$$h_t = o_t * \tanh(C_t) \tag{2.7}$$

where f_t is the forget gate, i_t is the input gate, o_t is the output gate, and C_t is the cell state.

Gated Recurrent Unit (GRU): GRUs [Cho et al., 2014] simplify LSTM by combining the forget and input gates into a single update gate:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{2.8}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{2.9}$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t]) \tag{2.10}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{2.11}$$

MinimalRNN: MinimalRNN is an even more parsimonious architecture designed to be interpretable while maintaining the ability to learn long-term dependencies.

2.3.2 Transformer Architecture

The Transformer architecture, introduced by Vaswani et al. (2017) [Vaswani et al., 2017], is based on the self-attention mechanism rather than recurrence. The key innovation is the multi-head attention mechanism, which allows the model to attend to different positions in the sequence simultaneously.
The scaled dot-product attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.12}$$

where Q (query), K (key), and V (value) are projections of the input sequence, and d_k is the dimension of the keys.
Multi-head attention concatenates multiple attention heads:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.13}$$

The Transformer encoder consists of alternating multi-head attention and feed-forward layers with residual connections and layer normalization.
Advantages of Transformers for temporal modeling include:

- **Parallel Processing**: Unlike RNNs, all positions in the sequence can be processed in parallel, enabling faster training.
- **Long-Range Dependencies**: Attention mechanisms can directly connect distant positions without intermediate steps, mitigating vanishing gradient problems.
- **Interpretability**: Attention weights provide insight into which parts of the input sequence the model considers most important.

2.3.3 Graph Neural Networks

Graph Neural Networks (GNNs) are designed to process data represented as graphs, where information is carried by nodes and edges. GNNs iterate message passing between nodes to learn node representations that incorporate information from their neighborhoods.

Graph Convolutional Networks (GCN): The basic GCN update rule performs convolution on a graph by aggregating information from neighboring nodes:

$$h_v^{(k+1)} = \sigma\left(W^{(k)} \sum_{u \in N(v)} \frac{1}{\sqrt{|N(u)||N(v)|}} h_u^{(k)}\right) \tag{2.14}$$

where $N(v)$ is the neighborhood of node v, $h_v^{(k)}$ is the representation of node v at layer k, and $W^{(k)}$ is a weight matrix.

GraphSAGE: GraphSAGE [Hamilton et al., 2017] extends GCN by explicitly sampling neighborhoods and using different aggregation functions:

$$h_v^{(k)} = \sigma\left(W^{(k)} \cdot \text{CONCAT}\left(h_v^{(k-1)}, \text{AGG}(\{h_u^{(k-1)} : u \in N(v)\})\right)\right) \tag{2.15}$$

where AGG can be mean, LSTM, or other aggregation functions.

Application to Brain Networks: Functional connectivity matrices naturally form graphs where brain regions are nodes and functional correlations are edges. GNNs can extract spatial patterns from these brain networks while preserving the network structure.

2.4 Neuroimaging Data

2.4.1 Structural MRI

Structural MRI provides high-resolution anatomical images of the brain. Key measures derived from structural MRI include:

- **Brain Volumes**: Whole brain volume, gray matter volume, white matter volume
- **Regional Atrophy**: Hippocampal volume, entorhinal cortex volume, ventricle volume

These volumetric measures are sensitive to neurodegeneration in AD, with the hippocampus being particularly vulnerable.

2.4.2 Resting-State Functional MRI

Resting-state fMRI (rs-fMRI) measures blood-oxygen-level-dependent (BOLD) signals while subjects rest without performing a specific task. From these signals, functional connectivity can be derived, reflecting temporal correlations between brain regions.

Functional Connectivity Matrix: A functional connectivity matrix is typically computed by: 1. Parcellating the brain into regions of interest (ROIs) using anatomical atlases 2. Extracting the BOLD time series from each ROI 3. Computing pairwise correlations between time series

Common parcellation schemes include:

- AAL (Automated Anatomical Labeling): 116 regions
- Schaefer Atlas: Available in multiple resolutions (100, 200, 400 regions)

- Power Atlas: 264 regions

Functional connectivity patterns have been shown to correlate with cognitive status and disease progression in AD.

2.4.3 PET Imaging

Positron emission tomography (PET) provides molecular imaging of AD pathology:

- **Amyloid PET**: Measures amyloid-beta burden using tracers like Florbetapir
- **Tau PET**: Measures tau pathology using various tau tracers
- **FDG-PET**: Measures glucose metabolism, which declines in AD

These biomarkers directly reflect AD pathology and are highly predictive of cognitive decline.

2.5 Longitudinal Data Challenges

Both approaches in this thesis address common challenges in longitudinal prediction:

1. **Irregular Visit Intervals**: Patients do not visit at regular intervals. Temporal information must be explicitly encoded to handle variable gaps.
2. **Missing Data**: Missing values occur at both the visit level (not all measurements at all visits) and subject level (some subjects have fewer visits). Imputation strategies are needed.

3. **Class Imbalance**: Stable subjects greatly outnumber converter subjects, leading to models biased toward predicting stability. Resampling and loss function modifications (e.g., focal loss) are used.

4. **Variable History Length**: Different subjects have different numbers of prior visits. Models must handle variable-length sequences.

5. **Information Decay**: Earlier visits may be less informative than recent visits. Models should learn which visits are most relevant.

The following chapters describe how the Transformer-based and GNN-based approaches address these challenges through different architectural and preprocessing strategies.

Chapter 3

Transformer-Based Longitudinal Prediction

This chapter presents the Transformer-based approach for predicting Alzheimer's disease progression. The model utilizes multi-modal clinical data from longitudinal visits to predict the diagnosis at a subject's next clinical visit. We describe the data, preprocessing, model architecture, training procedures, and comprehensive results.

3.1 Data and Materials

3.1.1 TADPOLE Challenge Dataset

We use data from the Alzheimer's Disease Prediction of Longitudinal Evolution (TADPOLE) challenge [Marinescu et al., 2018a], which contains longitudinal data from 1677 participants from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [Mueller et al., 2005]. TADPOLE provides longitudinal data

with multiple collection time points for each participant, with a minimum interval of 6 months between visits, making it ideal for longitudinal studies of disease progression.
The TADPOLE challenge specifically aimed to develop models to accurately predict future clinical states and biomarker trajectories in individuals at risk for or diagnosed with AD.

3.1.2 Features

We use a subset of 23 features from the TADPOLE dataset, the same ones recommended by the TADPOLE challenge organizers. These features span multiple data modalities:

- **Neuropsychological Tests** (10 features): FAQ, CDR-SB, MMSE, ADAS-Cog11, RAVLT immediate, RAVLT learning, ADAS-Cog13, RAVLT forgetting, RAVLT forgetting percent, MOCA
- **MRI Volumetric Measures** (7 features): Intracranial volume, whole brain volume, ventricles volume, hippocampus volume, entorhinal cortical volume, fusiform cortical volume, middle temporal cortical volume
- **PET Biomarkers** (2 features): Fluorodeoxyglucose (FDG) PET, Florbetapir PET
- **CSF Biomarkers** (4 features): Phosphorylated tau, beta-amyloid, total tau
- **Clinical Diagnosis**: CN, MCI, or AD

These features capture information about cognitive function, brain structure, and molecular AD pathology, providing a comprehensive view of disease state. The features show varying degrees of missingness, ranging from approximately 30% for cognitive scores to over 80% for CSF and PET biomarkers.

3.2 Preprocessing

3.2.1 Data Cleaning

The TADPOLE dataset, like many long-term studies, contains several quality issues that require preprocessing:

1. **Drop subjects with single visit**: We require each subject to have multiple visits to construct longitudinal sequences.

2. **Remove reverter subjects**: We identify and exclude subjects who convert to a less severe disease stage (AD to MCI, MCI to CN, or AD to CN). These reversions likely represent diagnostic errors rather than true disease reversions.

3. **Handle multiple conversions**: For subjects who convert from CN to MCI or MCI to AD multiple times, we retain only the first conversion and discard all visits after the initial conversion.

4. **Remove missing diagnoses**: We drop all visits with missing clinical diagnosis values, as the target variable is required.

3.2.2 Missing Data Imputation

The high degree of missingness in the dataset requires a principled imputation approach. Rather than discarding visits with missing features, we employ model filling, an imputation strategy that uses a minimalRNN model to predict missing values from observed data.

Model filling works as follows: A minimalRNN model is trained to predict each feature from other features and time information at each visit. The trained model then provides predictions for missing feature values that preserve the statistical properties of the observed data while leveraging correlations between features.

3.2.3 Temporal Feature Encoding

Since visits occur at irregular intervals, we introduce a temporal feature representing the number of months until the final visit. This feature encodes temporal information into every visit in a subject's visit history, allowing the model to account for variable time gaps between visits.

3.3 Sequence Generation and Dataset Construction

3.3.1 Converter and Stable Sequences

To focus on the key points of disease progression, we construct datasets using two types of sequences:

Definition 3.3.1 (Converter Sequence)**.** *A sequence of visits for a subject sorted by examination date where at the last visit in the sequence, the subject converts*

to the next stage of the disease (CN to MCI, CN to AD, or MCI to AD).

Definition 3.3.2 (Stable Sequence)**.** *A sequence of visits for a subject sorted by examination date where at the last visit in the sequence, the subject's diagnosis remains the same as the penultimate visit (CN to CN, MCI to MCI, or AD to AD).*

We generate one converter sequence of maximal length for every subject who converted. Similarly, we generate one stable sequence of maximal length from stable subjects. This creates the raw dataset containing 1222 stable sequences and 405 converter sequences (with only 70 containing CN to MCI conversions).

3.3.2 Sequence Grouping

Definition 3.3.3 (Group n)**.** *All stable or converter sequences with $n+1$ visits.*

Each sequence belongs to exactly one group based on its length. This grouping allows us to analyze model performance across different history lengths, which is clinically important as shorter histories may provide less information for prediction while longer histories may be noisier.

3.3.3 Dataset Balancing

The raw dataset is highly imbalanced, with far more stable than converter sequences. We create a balanced dataset by randomly downsampling stable sequences in each group until equal numbers of stable and converter sequences are present. This prevents models from being biased toward predicting stability and ensures equal learning from both classes.

Figure 3.1 shows the distribution of sequences by group in both raw and balanced datasets.

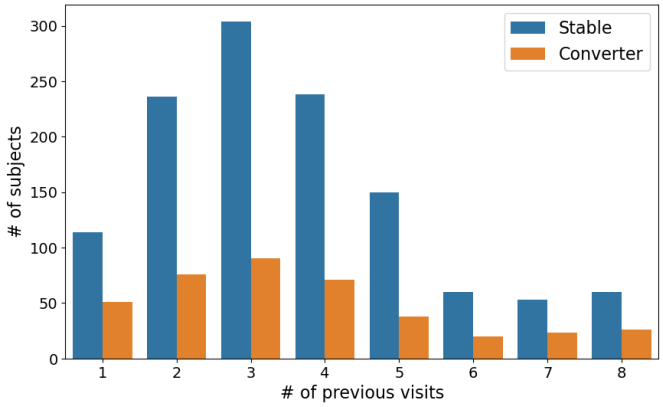

Figure 3.1: Distribution of visit sequences by group number in raw and balanced datasets

3.4 Model Architecture

3.4.1 Problem Formulation

We formulate this as a multi-class classification problem with three output classes (CN, MCI, AD). Given a sequence of n visits $v = (v_1, v_2, \ldots, v_n)$, where each v_i is a feature vector, the task is to predict the diagnosis d_{n+1} at visit $n + 1$.

3.4.2 Baseline RNN Models

We compare against three RNN-based baseline models:

1. **Long Short-Term Memory (LSTM)**: A standard deep learning baseline with memory cells for capturing long-term dependencies. We use a 2-layer LSTM with 128 hidden units.

2. **Gated Recurrent Unit (GRU)**: A simpler variant of LSTM with fewer parameters. We use a 2-layer GRU with 128 hidden units.

3. **MinimalRNN**: A parsimonious RNN architecture designed for interpretability while maintaining ability to learn long-term dependencies.

All RNN models take variable-length sequences as input, process them sequentially, and output a prediction based on the final hidden state.

3.4.3 Proposed Transformer Model

The proposed Transformer model processes the entire visit sequence using multi-head self-attention mechanisms rather than sequential processing. The architecture includes:

1. **Positional Encoding**: Since Transformers do not have inherent sequential bias like RNNs, we add positional encodings to each visit to preserve sequence order information.

2. **Multi-Head Self-Attention**: Multiple attention heads allow the model to attend to different aspects of the visit sequence. Each head learns different patterns of dependencies between visits.

3. **Feed-Forward Networks**: Each layer includes feed-forward networks with ReLU activation and residual connections.

4. **Layer Normalization**: Applied before each sub-layer for stable training.

5. **Global Average Pooling**: The output features are aggregated across all visits using global average pooling.

6. **Classification Head**: A linear layer maps the aggregated features to the three output classes.

3.4.4 Time Series Transformer (TST)

We also compare against a Time Series Transformer (TST) model optimized specifically for time series forecasting. The TST is an encoder-only Transformer that incorporates:

- **Causal Attention**: Attention masks that prevent the model from attending to future positions
- **Temporal Positional Encoding**: Positional encodings designed specifically for temporal data
- **Time-Series-Specific Inductive Biases**

3.5 Training Procedures

3.5.1 Hyperparameters

For all models, we use:

- **Optimizer**: Adam with learning rate 10^{-4}
- **Batch Size**: 32
- **Dropout Rate**: 0.3
- **Weight Decay**: 10^{-5}
- **Early Stopping**: Based on validation loss with patience of 50 epochs

- **Loss Function**: Cross-entropy loss

3.5.2 Cross-Validation

We use 10-fold cross-validation on both raw and balanced datasets. For each fold:

1. Data is split into training (70%), validation (15%), and testing (15%) sets
2. Models are trained on the training set with validation-based early stopping
3. Performance is evaluated on the held-out test set

Statistical significance of differences between models is assessed using two-sample Welch's t-tests with $n_1 = n_2 = 10$ (the number of folds).

3.5.3 Evaluation Metrics

We report multiple metrics to comprehensively evaluate model performance:

- **Accuracy**: Proportion of correctly classified instances
- **Balanced Accuracy (BCA)**: Average of sensitivity and specificity, less sensitive to class imbalance
- **Sensitivity**: True positive rate (ability to identify converters)
- **Specificity**: True negative rate (ability to identify stable cases)
- **Multi-class AUC (mAUC)**: Extension of AUC to multi-class setting
- **F1 Score**: Harmonic mean of precision and recall

For class-specific analysis, we evaluate performance separately on stable sequences and converter sequences, recognizing that these represent fundamentally different tasks of different clinical importance.

3.6 Results

3.6.1 Overall Performance Comparison

On the raw dataset, the Transformer model achieved the highest F1 score (0.824 ± 0.04) and demonstrated the highest overall accuracy (0.824 ± 0.04) compared to baseline RNN models. The Time Series Transformer (TST) achieved the highest mAUC (0.938 ± 0.01), though the standard Transformer was competitive (0.920 ± 0.02).

Table 3.1: Overall Performance Metrics (Raw Dataset)

Model	**Accuracy**	**BCA**	**mAUC**
TST	0.814 ± 0.04	0.866 ± 0.02	**0.938 ± 0.01**
Transformer	**0.824 ± 0.04**	**0.874 ± 0.03**	0.920 ± 0.02
GRU	0.792 ± 0.04	0.830 ± 0.05	0.846 ± 0.11
LSTM	0.789 ± 0.04	0.829 ± 0.06	0.857 ± 0.11
minRNN	0.777 ± 0.04	0.819 ± 0.06	0.873 ± 0.11

Statistical testing revealed that the Transformer significantly outperformed GRU (the best RNN) in Balanced Accuracy ($p = 0.041$), with the advantage primarily driven by significantly better sensitivity ($p = 0.039$). When considering raw accuracy, the Transformer outperformed minRNN ($p = 0.015$).

On the balanced dataset, the Transformer model again outperformed all RNN baselines across all metrics, with particularly strong performance in Balanced Accuracy (0.839 ± 0.04). Statistical testing confirmed the Transformer was significantly better than all RNNs; for the best RNN (LSTM), we observed significant differences in accuracy ($p = 0.0007$), BCA ($p = 0.0024$), sensitivity ($p = 0.0020$), and specificity ($p = 0.0052$).

3.6.2 Converter vs. Stable Sequence Performance

A key distinction emerges when examining performance separately on stable and converter sequences. This analysis reveals the fundamental difference between the model architectures.

Stable Sequence Performance

On stable sequences, the RNN models significantly outperformed the Transformer. For example, on the raw dataset:

- minRNN achieved 98.6% accuracy on stable sequences
- GRU achieved 98.4% accuracy
- The Transformer achieved only 88.3% accuracy

Statistical testing confirmed these differences were significant. The RNN models' superior performance on stable sequences suggests they have learned the dominant class bias well, essentially predicting stability with high confidence.

Converter Sequence Performance

On converter sequences, the situation reverses dramatically. The Transformer substantially outperformed all RNN models:
On the raw dataset converter sequences:

- Transformer achieved 69.4% accuracy with 0.662 BCA
- TST achieved 69.4% accuracy with 0.662 BCA
- GRU achieved only 18.4% accuracy with 0.464 BCA

- LSTM achieved 21.7% accuracy with 0.476 BCA
- minRNN achieved 12.9% accuracy with 0.442 BCA

This represents a dramatic advantage for the Transformer-based models, particularly for identifying disease conversion. Statistical testing confirmed TST's significantly higher accuracy ($p = 1.48 \times 10^{-10}$), BCA ($p = 1.01 \times 10^{-10}$), sensitivity ($p = 1.13 \times 10^{-6}$), and specificity ($p = 1.46 \times 10^{-10}$) compared to LSTM.

3.6.3 Effect of Visit History Length

Analysis of performance across different visit sequence lengths reveals important insights about how much prior history is needed for accurate prediction.
Both model types show similar overall trends across sequence lengths, but key distinctions emerge in the converter subset. The Transformer exhibits clear advantage in predicting conversion, particularly for sequences with shorter visit histories. Statistical analysis shows the Transformer significantly outperforms all RNN models in distinguishing converters in the earliest visit groups, with p-values reaching as low as 4.24×10^{-10} in group 1. This advantage remains significant through group 3 ($p = 1.04 \times 10^{-5}$) but becomes statistically insignificant for longer histories.
Interestingly, model performance does not consistently improve with longer visit sequences, contrary to initial expectations. Several factors likely contribute:

1. **Data Availability**: Groups 5-8 account for only 25% of the raw dataset
2. **Class Distribution**: Future diagnoses in groups 5-6 are disproportionately MCI, which is inherently more difficult to predict

3. **Temporal Structure**: Longer visit histories show larger intervals between visits (10.02 months on average vs 6.81), introducing more temporal inconsistency

4. **Feature Trends**: Linear regression analysis shows that longer sequences have weaker overall trends, with smaller regression slopes indicating less pronounced change over time

Despite these challenges, ablation studies confirm that the model leverages information from longer visit histories, as performance degrades when fewer visits are provided. However, the decrease is modest, emphasizing the importance of recent visits in making accurate diagnoses—an expectation consistent with clinical intuition.

3.6.4 Feature Importance Analysis

Feature ablation studies reveal the relative importance of different data modalities. Results show:

- **Cognitive Scores**: Removing cognitive test scores decreased BCA by 15.76%, indicating these provide the most information
- **Volumetric MRI**: Removing volumetric MRI features resulted in relatively modest performance changes
- **Biomarkers**: Removing PET and CSF biomarkers showed minimal impact, likely due to their high missingness

When used in isolation, cognitive scores alone achieved nearly the same BCA as the full feature set, while volumetric MRI and biomarker features alone showed greater performance declines.

Analysis of misclassified sequences reinforces the importance of cognitive scores. Many errors occurred when cognitive scores were inconsistent with the overall class average. For instance, AD sequences misclassified as MCI showed substantially lower ADAS13 scores in the penultimate visit compared to typical AD visits. This difference was statistically significant ($p = 1.89 \times 10^{-19}$ on Mann-Whitney U test), suggesting the model struggles to account for individual variability in cognitive test performance.

3.7 Discussion

The results demonstrate that the Transformer model significantly outperforms RNN-based models in identifying converter sequences, which is critical for early intervention in AD. This improvement comes at the cost of slightly lower performance on stable sequences, but this trade-off is clinically favorable since false negatives (missing converters) carry higher clinical consequence than false positives (predicting conversion when stability occurs).

The superior performance of Transformers on converter identification, particularly for shorter visit histories, suggests that attention-based architectures are better suited for capturing the subtle patterns that distinguish early disease transitions. The RNN models appear to have learned the dataset imbalance by confidently predicting stability, a learned bias that is difficult to overcome even with balanced datasets.

The finding that recent visits are more informative than older visits is consistent with clinical expectations. Recent visits provide more current information about disease state and allow assessment of recent rate of change, which are more predictive of near-term progression than historical information.

Chapter 4

Graph Neural Network Approach with Functional Connectivity

This chapter presents the history-aware graph neural network (HA-GNN) approach [Moghaddami et al.] for predicting cognitive impairment progression using functional connectivity from resting-state fMRI. Unlike the Transformer approach in Chapter 3 which operates on structured tabular features, this approach preserves the spatial brain network structure inherent in functional connectivity matrices, using Graph Neural Networks to extract meaningful patterns.

4.1 Data and Materials

4.1.1 ADNI Dataset

We use resting-state functional MRI (rs-fMRI) and structural MRI (sMRI) data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [Jack et al., 2008],

supplemented with diagnosis information and longitudinal visit data from the TADPOLE challenge [Marinescu et al., 2018b]. The TADPOLE dataset provides longitudinal tracking information for 1677 participants, from which we select subjects who meet specific criteria:

1. At least two rs-fMRI scans (to enable longitudinal prediction)
2. At least one structural MRI scan (for preprocessing and brain atlas alignment)
3. Longitudinal diagnosis data available

After applying these criteria and removing "reverter" subjects (those transitioning to less severe cognitive stages), we obtain 303 participants, consisting of:

- 250 stable subjects (no change in diagnosis at last visit)
- 53 converter subjects (transitioning to more severe stage at last visit)

The 303 subjects provide 1089 total rs-fMRI scans, with 894 scans from stable subjects and 195 scans from converter subjects. Visit intervals are irregular with a mean of 14.78 months between consecutive scans. Subjects with longer visit histories (>7 scans) are rare.

4.1.2 Converter and Stable Subject Definitions

Definition 4.1.1 (Converter Subject)**.** *A subject that transitions to a more severe stage of cognitive impairment (CN to MCI, CN to AD, or MCI to AD) in their last visit compared to the penultimate visit.*

Definition 4.1.2 (Stable Subject)**.** *A subject with the same diagnosis in their last visit as their penultimate visit (CN to CN, MCI to MCI, or AD to AD).*

Notably, there are no subjects converting from CN directly to AD in their final visit, making the CN to MCI transition the earliest conversion point and most clinically relevant for early intervention.

4.2 Preprocessing

4.2.1 fMRIPrep Pipeline

Raw fMRI data requires extensive preprocessing to remove artifacts and prepare for analysis. We apply the fMRIPrep preprocessing pipeline with default parameters, which includes:

1. **Anatomical Preprocessing**: Skull stripping to remove non-brain tissue
2. **Brain Tissue Segmentation**: Classification into cerebrospinal fluid (CSF), white matter (WM), and gray matter (GM)
3. **Spatial Normalization**: Alignment to MNI152NLin2009cAsym standard space for group analysis
4. **Functional Preprocessing**:
    - Head motion correction using volume-to-volume rigid body registration
    - Slice-timing correction for interleaved acquisition
    - Co-registration of functional to structural images

5. **Confound Regression**: Extraction of motion parameters and other confound time series

6. **Resampling**: Output to target space with standard resolution

Raw DICOM files are first converted to NIfTI format and organized in Brain Imaging Data Structure (BIDS) format before applying the fMRIPrep pipeline.

4.3 Functional Connectivity Matrix Computation

4.3.1 Brain Parcellation

Rather than analyzing individual voxels, we parcellate the brain into regions of interest (ROIs) to reduce dimensionality while preserving functionally meaningful brain regions. We use the Schaefer atlas to parcellate the brain into 100 ROIs. Each ROI represents a functionally coherent region with similar connectivity patterns. The parcellation provides a natural structure for constructing functional connectivity networks.

4.3.2 Time Series Extraction and Correlation Computation

For each ROI, we extract the BOLD time series by averaging signals from all voxels within that region. The functional connectivity matrix is computed as the pairwise Pearson correlations between the 100 ROI time series:

$$FC_{ij} = \text{corr}(\text{BOLD}_i, \text{BOLD}_j) \tag{4.1}$$

where $i, j \in \{1, \ldots, 100\}$ are ROI indices. The resulting FC matrix is a 100×100 symmetric matrix where element (i, j) represents the functional correlation between regions i and j.

The functional connectivity matrix naturally forms a graph structure suitable for graph neural network processing, where ROIs are nodes and correlations are edge weights.

4.4 Model Architecture

4.4.1 Problem Formulation

We formulate this as a binary classification task: predicting whether a subject will convert to a more severe stage of cognitive impairment at their last visit. Given a subject's visit history $v = (v_1, v_2, \ldots, v_{n+1})$ where each v_i is the functional connectivity matrix from the i-th visit, we predict whether $d_n = d_{n+1}$ (stable) or $d_n \neq d_{n+1}$ (converter), where $d_i \in \{\text{CN}, \text{MCI}, \text{AD}\}$ is the cognitive diagnosis at visit i.

4.4.2 Overall Architecture

The HA-GNN model combines two main components:

1. **Graph Convolutional Network (GCN)**: Processes spatial brain network information within each visit's functional connectivity matrix
2. **Recurrent Neural Network (RNN)**: Models temporal evolution across visits
3. **Temporal Encoding**: Explicit encoding of time intervals between visits

Figure 4.1 shows the complete architecture.

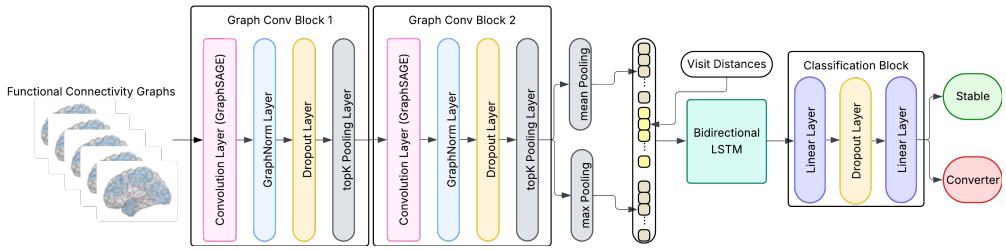

Figure 4.1: Architecture of the HA-GNN model showing GCN processing of functional connectivity graphs followed by RNN for temporal modeling

4.4.3 Graph Convolutional Network Component

The GCN component is loosely based on the BrainGNN architecture [Li et al., 2020] but adapted for processing multiple visits. Each GCN block consists of:

1. **GraphSAGE Layer**: Performs neighborhood aggregation to learn node representations that incorporate information from neighboring ROIs.

   The GraphSAGE update is:

$$h_v^{(k)} = \sigma\left(W^{(k)} \cdot \text{CONCAT}\left(h_v^{(k-1)}, \text{AGG}(\{h_u^{(k-1)} : u \in N(v)\})\right)\right) \tag{4.2}$$

   where h_v is the representation of node (ROI) v, $N(v)$ is its neighborhood, and AGG is an aggregation function (typically mean).

2. **Graph Normalization Layer**: Normalizes features across the graph for more stable learning

3. **Dropout Layer**: Regularization to prevent overfitting

4. **TopK Pooling Layer**: Selects the most important nodes (ROIs) based on learned importances, reducing dimensionality while preserving the most discriminative brain regions

The GCN has two convolutional blocks, each followed by pooling, progressively capturing more abstract patterns of brain connectivity.

4.4.4 Temporal Feature Encoding

Between the GCN and RNN components, we concatenate the distances (in months) between consecutive visits. This provides the RNN with explicit temporal information about:

- How far in the past each visit occurred
- How much time has elapsed between consecutive observations
- The time horizon for the prediction

This explicit temporal encoding helps the model account for irregular visit spacing and calibrate predictions based on the time interval.

4.4.5 Recurrent Neural Network Component

We evaluate two RNN architectures for the temporal component:

1. **Long Short-Term Memory (LSTM)**: Uses memory cells with explicit gates for learning what to store, forget, and output. LSTMs excel at capturing long-term dependencies and are particularly suitable for the multi-month intervals in our data.

2. **Gated Recurrent Unit (GRU)**: A simplified RNN variant with fewer parameters than LSTM, combining forget and input gates into a single update gate.

The RNN processes the sequence of GCN features, integrating temporal information about disease progression.

4.4.6 Prediction Head

The RNN output is passed through a fully connected layer that outputs a probability for the binary classification (stable vs. converter).

4.5 Training and Evaluation

4.5.1 Two-Stage Training Strategy

Stage 1: GCN Pretraining

Before training the full HA-GNN model, we pretrain the GCN component on a 3-class classification task: predicting a subject's diagnosis (CN, MCI, or AD) from a single visit's functional connectivity matrix.
This pretraining serves several purposes:

- Provides meaningful initialization of the GCN weights
- Helps the GCN learn representations of functional connectivity patterns associated with different disease states
- Reduces training time for the full model

The pretraining uses a random 20% split of the data and trains only the GCN component without the RNN.

Stage 2: Full Model Training

After pretraining, we use 5-fold cross-validation to train and evaluate the complete HA-GNN model on the longitudinal prediction task. The GCN weights are initialized from pretraining and fine-tuned during full model training.

4.5.2 Hyperparameter Tuning

We employ Bayesian Optimization for hyperparameter tuning, which is more efficient than grid search for high-dimensional hyperparameter spaces. This automated approach finds optimal settings for:

- Learning rate
- Dropout rates
- Number of attention heads (if applicable)
- Other architectural parameters

4.5.3 Addressing Class Imbalance

The dataset is substantially imbalanced (250 stable vs. 53 converter subjects). We address this through:

1. **Focal Loss**: A variant of cross-entropy loss that down-weights easy examples and focuses training on hard examples. Focal loss is defined as:

$$\mathrm{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \tag{4.3}$$

where p_t is the probability of the true class and γ is a focusing parameter. We use $\alpha = 0.9, \gamma = 3$.

2. **Balanced Accuracy Metric**: During validation and testing, we use balanced accuracy (average of sensitivity and specificity) rather than raw accuracy, which is more appropriate for imbalanced data.

4.6 Results

4.6.1 Overall Performance

The HA-GNN model with LSTM RNN component achieved the best overall performance:

Table 4.1: Classification results of different HA-GNN variants

Model	**Acc.**	**AUC**	**BAC**
HA-GNN (LSTM)	**82.9 ± 5.8**%	**85.2 ± 6.5**%	**77.1 ± 11.4**%
HA-GNN (GRU)	73.3 ± 4.9%	78.6 ± 6.2%	70.4 ± 7.1%
HA-GNN (RNN)	74.3 ± 3.8%	77.1 ± 6.7%	65.1 ± 11.4%
w/o Pretraining	72.3 ± 7.9%	64.2 ± 9.5%	59.4 ± 6.1%

Table 4.2: Conversion results of different HA-GNN variants

Model	**CN to MCI**	**MCI to AD**
HA-GNN (LSTM)	**68.8**%	**67.6**%
HA-GNN (GRU)	62.5%	67.6%
HA-GNN (RNN)	50.0%	51.4%
w/o Pretraining	37.5%	40.5%

The LSTM variant substantially outperforms both GRU and vanilla RNN variants, with balanced accuracy of 77.1% ± 11.4% and overall accuracy of 82.9% ± 5.8%. The performance advantage of LSTM is consistent with its

known advantages in capturing long-term dependencies, particularly important given the multi-month intervals between visits.

The high standard deviation in balanced accuracy (11.4%) reflects sensitivity to the particular distribution of converter subjects in each cross-validation fold. This is expected given the substantial class imbalance and highlights the importance of focal loss in ensuring consistent identification of the minority converter class.

4.6.2 Early Conversion Detection

A particularly notable result is the model's strong performance on CN to MCI conversion: 68.8% accuracy. This early-stage transition is clinically most relevant for intervention, as it represents the earliest point at which disease progression can be detected. Achieving 68.8% accuracy at this early transition demonstrates that functional connectivity patterns contain meaningful information about incipient cognitive decline.

The model also achieves competitive performance on MCI to AD conversion (67.6% accuracy).

4.6.3 Impact of GCN Pretraining

The importance of GCN pretraining is demonstrated by the performance degradation when this step is omitted. Without pretraining:

- Accuracy drops from 82.9% to 72.3%
- Balanced Accuracy drops from 77.1% to 59.4%
- CN to MCI accuracy drops from 68.8% to 37.5%

This substantial degradation (drop of 30+ percentage points on CN to MCI detection) demonstrates that pretraining provides crucial initialization that enables the model to learn meaningful representations of disease-related functional connectivity patterns.

4.6.4 Comparison with Existing Approaches

While direct comparison with existing literature is challenging due to the novelty of the methodology, we can contextualize our results:

- **Kim et al.**: Investigated specific ROIs and their effectiveness in detecting different cognitive stages, providing complementary interpretability analysis.
- **Grammenos et al.**: Focused on distinguishing stable MCI (sMCI) from progressive MCI (pMCI) patients using XGBoost, achieving 86% accuracy. However, their more narrowly defined task and use of engineered features (rather than raw imaging data) make direct comparison difficult.
- **Wang et al.**: Used rs-fMRI from ADNI to identify AD subjects, achieving 67.65% accuracy. Our 82.9% accuracy on the related task of predicting conversion represents competitive performance.

Our approach operates directly on functional connectivity matrices, preserving rich spatial information in brain network topology, which differentiates it from feature-engineering approaches.

4.7 Discussion

4.7.1 Architecture Advantages

The HA-GNN approach demonstrates several key advantages:

1. **Spatial-Temporal Integration**: By combining GCN for spatial patterns and RNN for temporal evolution, the model captures both organizational characteristics of functional brain networks and their changes over time.

2. **Handling Irregular Spacing**: Explicit encoding of visit intervals allows the model to appropriately weight temporal information despite irregular visit spacing.

3. **End-to-End Learning**: Rather than relying on hand-crafted ROI features, the model learns which brain regions and connections are most predictive.

4. **Robustness**: Achieves strong performance despite missing visits and incomplete visit histories.

4.7.2 LSTM vs. Other RNN Variants

The substantial superiority of LSTM (77.1% BA) over GRU (70.4% BA) and vanilla RNN (65.1% BA) reflects the advantages of LSTM's memory cell architecture in capturing long-term dependencies. Given that visits span months or years, LSTM's explicit memory mechanisms are particularly valuable for maintaining information across large temporal gaps.

4.7.3 Class Imbalance and Performance Variability

The high standard deviation in balanced accuracy (11.4%) across cross-validation folds indicates sensitivity to the distribution of converter subjects in each fold. This is expected given:

- Severe class imbalance (250 stable vs. 53 converter subjects)
- Small number of converter subjects overall
- Additional imbalance in conversion types (no CN to AD conversions)

The use of focal loss addresses but does not completely eliminate this variability. Future work might explore stratified cross-validation or other resampling strategies to better control fold composition.

4.7.4 Early-Stage Detection Significance

The 68.8% accuracy on CN to MCI detection is clinically significant. This stage represents the earliest and most favorable point for intervention, as significant neurodegeneration has not yet occurred. Achieving nearly 69% accuracy at this critical transition, despite it being the most difficult to detect, demonstrates meaningful signal in functional connectivity patterns.

4.7.5 Information Encoded in Functional Connectivity

The model's ability to predict cognitive transitions from functional connectivity alone suggests that rs-fMRI-derived connectivity patterns capture important information about cognitive status and disease progression. These patterns likely reflect:

- Network reorganization in response to incipient neurodegeneration
- Disruptions in large-scale brain networks known to be affected by AD (default mode network, etc.)
- Compensatory network changes in response to pathology

The high standard deviation in balanced accuracy suggests that while functional connectivity contains signal, considerable noise or individual variability remains. Future work incorporating additional data modalities or more sophisticated preprocessing may improve performance.

4.7.6 Comparison with Transformer Approach

While the HA-GNN approach achieves strong accuracy (82.9%), it is slightly lower than the Transformer approach on multi-class diagnosis prediction (82.4% on the balanced dataset). However, the HA-GNN operates on a fundamentally different task (binary conversion prediction from imaging) compared to the Transformer (multi-class diagnosis prediction from structured features). Direct comparison is therefore not straightforward, and the two approaches are best viewed as complementary.

Chapter 5

Comparative Analysis and Discussion

This chapter synthesizes findings across the two complementary approaches presented in Chapters 3 and 4, comparing their methods, strengths, limitations, and clinical implications. Understanding the complementary nature of these approaches provides insights into how to effectively predict Alzheimer's disease progression.

5.1 Overview of the Two Approaches

5.1.1 Transformer-Based Approach

Data Modality: Structured clinical and imaging-derived features (cognitive scores, structural MRI volumes, PET biomarkers, CSF markers)

Task: Multi-class diagnosis prediction (CN, MCI, AD) at next clinical visit

Data Scale: 1677 subjects with varying numbers of visits

Architecture: Transformer encoder with multi-head self-attention

Handling of Irregular Visits: Temporal features encoding months until final visit

Missing Data Strategy: Model filling using minimalRNN predictions

Key Result: 82.4% accuracy, 0.920 mAUC on balanced dataset; particularly strong converter identification

5.1.2 GNN-Based Approach

Data Modality: Functional connectivity matrices from resting-state fMRI

Task: Binary conversion prediction from functional connectivity

Data Scale: 303 subjects with 1089 rs-fMRI scans

Architecture: Graph Convolutional Network (spatial) + RNN (temporal)

Handling of Irregular Visits: Explicit encoding of temporal intervals between visits

Missing Data Strategy: Requires complete functional connectivity matrices; handles irregular visit histories naturally

Key Result: 82.9% overall accuracy, 77.1% balanced accuracy, 68.8% CN to MCI accuracy

5.2 Methodological Comparison

5.2.1 Feature Representation

Transformer Approach: Structured Features

The Transformer approach operates on manually curated features:

- Cognitive test scores directly quantify cognitive function
- Structural MRI volumes are derived biomarkers of neurodegeneration
- PET biomarkers directly measure AD pathology (amyloid, tau)
- CSF markers reflect pathological processes

These features are human-interpretable and have established clinical relevance. However, they require careful feature engineering, subject to availability of specific tests and imaging modalities at each visit. High missingness in some features (>80% for PET and CSF) creates challenges.

GNN Approach: Network-Structured Data

The GNN approach operates directly on functional connectivity matrices:

- Preserves network structure reflecting brain organization
- Allows learning of network patterns without predetermined feature engineering
- Weights learned by GraphSAGE reflect which brain regions and connections are most important
- Completely avoids missing feature problems (complete FC matrix available when fMRI scan is obtained)

The trade-off is that interpretability requires additional analysis to understand which brain regions drive predictions. The approach also requires that functional connectivity be computed consistently across subjects.

5.2.2 Temporal Modeling

Transformer: Attention-Based Temporal Dependencies

The Transformer processes the entire visit sequence in parallel through self-attention mechanisms. Multi-head attention allows the model to:

- Attend to different aspects of the sequence simultaneously
- Learn which visits are most relevant for prediction
- Capture dependencies between visits without the sequential bottleneck of RNNs
- Benefit from parallel processing, enabling faster training

Temporal information is encoded as additional features (months until final visit) rather than through the architectural structure.

GNN: RNN-Based Temporal Modeling

The GNN uses RNNs (particularly LSTM) to model temporal evolution:

- Sequential processing naturally aligns with temporal structure
- LSTM's memory cell explicitly maintains information across time steps
- Well-suited for variable-length sequences
- Temporal intervals are explicitly concatenated as features

The combination of GraphSAGE for spatial patterns and LSTM for temporal patterns creates a modular architecture capturing complementary information.

5.2.3 Handling of Dataset Challenges

Irregular Visit Spacing

Transformer: Encodes temporal information as features. The number of months until the final visit is added as a feature to each visit. Attention mechanisms then learn which visits are most relevant given temporal context.
GNN: Explicitly concatenates visit intervals as features to the RNN. The model can learn how prediction confidence should change based on elapsed time and distance to prediction.
Both approaches explicitly represent temporal information, though using different mechanisms. Neither approach requires visits to be equally spaced.

Missing Data

Transformer: Handles missing feature values through model filling—training a minimalRNN to predict missing values from observed features. This approach requires that clinical diagnosis is always available (as it's the target), but handles missing measurements for other features.
GNN: Avoids the problem by operating on complete functional connectivity matrices. When a subject has an rs-fMRI scan, a complete FC matrix is available; if a scan is missing, that visit is absent from the sequence but doesn't introduce partial data.
The GNN approach is simpler in this respect, avoiding imputation entirely at the cost of requiring complete imaging data when used.

Variable-Length Sequences

Transformer: Sequences of different lengths are processed through padding or masking. The self-attention mechanism can learn to attend only to real visits, ignoring padding. No explicit architecture changes needed.

GNN: RNNs naturally handle variable-length sequences through the sequential processing structure. Padding or masking is applied during preprocessing.

Both approaches handle variable-length sequences effectively through standard techniques.

Class Imbalance

Transformer: Addresses imbalance through dataset balancing (randomly downsampling stable sequences to match converter sequences in each visit-history-length group). This balances both the overall dataset and the distribution across history lengths.

GNN: Addresses imbalance through focal loss, a loss function that down-weights easy examples and focuses training on hard examples. No resampling is performed.

The Transformer's balancing approach trades training data for explicit 50-50 class balance. The GNN's focal loss approach preserves all training data while adjusting training dynamics. Each has trade-offs: balancing simplifies training but reduces training data; focal loss maintains data but requires careful tuning of loss parameters.

5.3 Performance Comparison

5.3.1 Accuracy Metrics

Table 5.1: Summary of key performance metrics for both approaches

Metric	**Transformer**	**GNN**
Overall Accuracy	82.4% (balanced dataset)	82.9%
Balanced Accuracy	87.4% (raw), 83.9% (balanced)	77.1%
mAUC / AUC-ROC	0.920 (balanced dataset)	0.852
Converter Detection	64.8% (raw), 79.2% (balanced)	69.4% (converter accuracy)

Both approaches achieve competitive overall accuracy (82-83%). However, the metrics reflect different tasks and evaluation protocols, making direct comparison complex.

5.3.2 Task Differences

A crucial distinction is that the two approaches address different prediction problems:

Transformer Task: Multi-class diagnosis prediction (CN vs. MCI vs. AD)

- Predicts specific cognitive stage at next visit
- 3 possible outputs
- Evaluated on both stable and converter sequences
- Reports class-wise performance metrics

GNN Task: Binary conversion prediction

- Predicts whether conversion occurs (yes/no)

- 2 possible outputs
- Evaluated separately for CN to MCI and MCI to AD transitions
- Optimized to distinguish stable from converter subjects

The Transformer must distinguish among three states, while the GNN only distinguishes two. This makes direct accuracy comparison challenging.

5.3.3 Early Detection Performance

A particularly clinically relevant metric is performance on early-stage detection:

Transformer:

- Specifically analyzes performance on sequences with shorter visit histories
- Achieves highest advantage on the earliest groups
- Demonstrates significant improvement ($p = 4.24 \times 10^{-10}$) over RNNs on shortest history group

GNN:

- Achieves 68.8% accuracy specifically on CN to MCI conversions
- This early transition is the most difficult to detect and most clinically relevant
- Strong performance on this most challenging transition is notable

Both approaches demonstrate particular strength in early detection, though through different metrics. The Transformer shows relative advantage for early time points when prior history is shortest. The GNN shows absolute accuracy on the CN to MCI transition, the cognitively earliest stage.

5.3.4 Converter Sequence Performance

Transformer Converter Performance

On the balanced dataset (equal numbers of stable and converter sequences):

- Overall accuracy on converters: 79.2%
- Balanced accuracy: 71.3%
- Sensitivity: 50.4%
- Specificity: 92.2%

Performance is strong but with lower sensitivity (true positive rate for identifying converters). This reflects a trade-off inherent in the multi-class setting where the model must distinguish among three diagnosis classes.

GNN Converter Performance

On converter subjects:

- Binary classification accuracy (convert vs. stable): High relative to baseline
- CN to MCI: 68.8%
- MCI to AD: 67.6%

The GNN's architecture specifically optimizes for binary conversion detection, resulting in more focused performance on this task.

5.4 Strengths and Limitations

5.4.1 Transformer Approach: Strengths

1. **Comprehensive Feature Coverage**: Integrates cognitive, structural, and molecular biomarkers in a single model

2. **Larger Sample Size**: Leverages 1677 subjects from TADPOLE, providing more training data

3. **Multi-Class Prediction**: Predicts specific cognitive stage, not just conversion

4. **Fast Training**: Parallel processing in Transformers enables efficient training compared to sequential RNNs

5. **Attention Interpretability**: Attention weights show which visits the model considers important

6. **Strong Converter Detection**: Particularly effective at identifying converter sequences

7. **History Length Analysis**: Systematic analysis of performance across different history lengths provides insights into which time scales matter most

5.4.2 Transformer Approach: Limitations

1. **Feature Engineering Dependent**: Relies on curated features that may not capture all relevant information

2. **Missing Data Issues**: Requires imputation strategy to handle features that are not collected at all visits

3. **Limited Feature Interpretability**: Unclear which specific cognitive scores or biomarkers drive individual predictions

4. **Modest Improvement Over RNNs**: While statistically significant, performance gains are relatively modest (82.4% vs. 79% for GRU)

5. **History Length Effect**: Performance does not improve consistently with longer visit histories, suggesting diminishing value of older information

6. **Data Modality Limitation**: Cannot directly incorporate imaging-derived network structure

5.4.3 GNN Approach: Strengths

1. **Network Preservation**: Maintains brain network structure inherent in functional connectivity

2. **No Feature Imputation**: Complete functional connectivity matrices eliminate missing data imputation needs

3. **End-to-End Learning**: Learns which brain regions and connections are important without predetermined feature selection

4. **Early Detection Performance**: Strong 68.8% accuracy on CN to MCI, the earliest and most clinically relevant transition

5. **Natural Handling of Irregular Visits**: Explicit temporal encoding naturally accommodates variable visit intervals

6. **Neurobiologically Grounded**: Operates directly on brain network structure, aligning with neuroscience understanding

7. **Complementary Data Modality**: Uses fMRI connectivity instead of tabular features, potentially capturing different pathological processes

5.4.4 GNN Approach: Limitations

1. **Smaller Sample Size**: Only 303 subjects with rs-fMRI data vs. 1677 in TADPOLE

2. **fMRI Data Requirements**: Requires complete rs-fMRI scans; cannot operate on subjects without imaging

3. **Binary Classification**: Only predicts conversion, not specific cognitive stage transition

4. **High Performance Variability**: Standard deviation of 11.4% in balanced accuracy reflects sensitivity to class distribution

5. **Interpretability Challenges**: Which brain regions drive predictions requires additional analysis

6. **Reduced Transparency**: Network pooling operations and learned aggregations obscure specific connectivity patterns used

7. **Preprocessing Dependencies**: Results depend on fMRI preprocessing pipeline choices (atlas selection, correlation threshold, etc.)

5.5 Complementary Nature

Rather than viewing these approaches as competing, they are best understood as complementary:

5.5.1 Different Data Modalities

The Transformer operates on tabular clinical features while the GNN operates on brain network connectivity. These capture different aspects of disease:

- **Clinical Features**: Direct measures of cognitive function (MMSE, ADAS-Cog) and structural outcomes (atrophy)
- **Functional Connectivity**: Indirect markers of functional brain organization and network disruption

Potentially, an ensemble approach combining both could leverage the complementary information.

5.5.2 Different Task Focuses

The Transformer's multi-class prediction provides more nuanced information (which stage will the patient reach?), while the GNN's binary conversion focus is sharper for the critical binary decision (will the patient convert?).
For clinical applications:

- Multi-class information aids treatment planning (knowing whether MCI or AD expected)
- Binary conversion information is critical for triage (who needs aggressive intervention?)

5.5.3 Different History Requirements

The Transformer shows that information from shorter histories is most valuable, particularly for early detection. The GNN, which also works with variable-length histories, could potentially benefit from the insight that recent visits are more informative.

5.5.4 Different Trade-offs

- **Accuracy vs. Completeness**: Transformer achieves higher overall accuracy; GNN achieves specialized accuracy on CN-to-MCI
- **Simplicity vs. Neurobiological Fidelity**: Transformer is simpler (tabular data); GNN preserves brain network structure
- **Training Data vs. No Imputation**: Transformer uses more data; GNN avoids imputation complexity
- **Attention Transparency vs. Feature Coverage**: Transformer provides attention visualizations; GNN integrates imaging-derived connectivity

5.6 Clinical Implications

5.6.1 Model Selection for Clinical Application

For clinical deployment, model selection depends on available data and clinical question:

When to Use Transformer Model:

- Cognitive test scores and MRI available but fMRI not available

- Clinician wants specific diagnosis prediction (CN, MCI, or AD stage)
- Maximizing accuracy on standard clinical measures is priority
- Larger patient populations with varying data availability

When to Use GNN Model:

- Research centers with available fMRI data
- Focus on binary high-risk identification (will convert: yes/no)
- Particular interest in functional brain network biomarkers
- Patients with missing structural or molecular biomarkers

5.6.2 Early Detection Priority

Both approaches demonstrate that early detection is possible with automated methods. The Transformer's focus on shortest history groups and the GNN's strong CN to MCI performance both suggest that meaningful signals emerge even with limited longitudinal data.
However, challenges remain:

- GNN achieves only 68.8% on CN to MCI despite being the easiest structural transition
- Transformer shows diminishing returns with longer histories despite initial expectation of improvement
- Both approaches show high baseline error rates, indicating that disease progression prediction remains genuinely difficult

These results should manage clinical expectations: automated prediction can meaningfully improve over chance but cannot achieve certainty in individual patient prediction.

5.6.3 False Positive vs. False Negative Trade-offs

Both approaches reveal important trade-offs:

Transformer: Slightly lower stable sequence performance in exchange for much better converter identification. This is clinically favorable since missing a converter (false negative) has higher clinical cost than incorrectly predicting conversion (false positive).

GNN: High specificity (92.2%) with moderate sensitivity reflects conservative prediction of conversion. This may appropriately reflect the inherent difficulty of early prediction before pathology becomes unambiguous.

For clinical use, stakeholders must decide acceptable false positive and false negative rates based on available interventions and their side effects.

5.6.4 Longitudinal Monitoring Value

Both approaches demonstrate that longitudinal information is valuable. The Transformer's finding that recent visits matter most and the GNN's ability to incorporate irregular histories both support the value of repeated clinical visits, even when spacing is irregular.

Clinically, this supports continued longitudinal monitoring rather than single-timepoint assessment. The finding that early groups show better model performance suggests particular value in frequent monitoring at early disease stages.

5.7 Future Directions for Synthesis

5.7.1 Ensemble Approaches

Given the complementary nature, ensemble methods combining both approaches could potentially leverage their respective strengths:

- Use Transformer predictions as clinical features
- Combine outputs through learned weights
- Ensemble decision making for highest confidence predictions

5.7.2 Multi-Modal Integration

Rather than treating approaches separately, developing unified models that jointly process:

- Tabular clinical features
- Functional connectivity networks
- Potentially structural imaging features

could leverage all available information.

5.7.3 Biomarker-Guided Model Selection

Different patients have different biomarker profiles (some with CSF data, some with fMRI, etc.). Adaptive model selection based on available data could optimize performance for each patient.

5.7.4 Temporal Forecasting

Both models predict the next clinical visit. Extensions could:

- Predict multiple steps into the future
- Estimate time to conversion
- Provide uncertainty estimates for probabilistic clinical decision support

Chapter 6

Conclusion and Future Work

6.1 Summary of Findings

This thesis presented two complementary deep learning approaches for predicting Alzheimer's disease progression using longitudinal data. The research demonstrates that modern deep learning architectures can effectively address the complex challenge of forecasting cognitive decline in individuals at risk for or diagnosed with Alzheimer's disease.

6.1.1 Transformer-Based Approach

The Transformer-based model for multi-class diagnosis prediction achieved 82.4% accuracy and 0.920 mAUC on the balanced TADPOLE dataset. Key findings include:

1. **Superior Converter Detection**: The Transformer significantly outperformed RNN baselines (LSTM, GRU, minimalRNN) in identifying converter sequences, achieving 79.2% accuracy on converters in the

balanced dataset. This improvement was statistically significant ($p < 0.05$).

2. **Early Detection Advantage**: The model showed particular advantage for sequences with shorter visit histories, achieving statistical significance ($p = 4.24 \times 10^{-10}$) over RNNs on the shortest history group. This early detection capability is clinically valuable.

3. **Feature Importance**: Cognitive test scores (FAQ, MMSE, ADAS-Cog, etc.) provided the most information for prediction, with 15.76% BCA decrease when removed. Volumetric MRI and biomarkers contributed less directly.

4. **Recent Visit Primacy**: The model demonstrated that recent visits are more informative than distant historical data. While longer histories helped, the improvements were modest, and performance did not consistently improve with sequence length.

5. **Attention Mechanism Value**: The parallel attention-based processing of the Transformer proved more effective than sequential RNN processing for this task, particularly for identifying subtle patterns distinguishing converters from stables.

6.1.2 Graph Neural Network Approach

The HA-GNN model for binary conversion prediction from functional connectivity achieved 82.9% overall accuracy and 77.1% balanced accuracy. Key findings include:

1. **Early Transition Detection**: Achieved 68.8% accuracy specifically on CN to MCI conversions, the earliest and most clinically relevant cognitive

transition. This demonstrates that functional connectivity patterns contain meaningful information about incipient cognitive decline.

2. **LSTM Superiority**: Among RNN variants tested, LSTM substantially outperformed GRU (77.1% vs. 70.4% balanced accuracy) and vanilla RNN (65.1%), reflecting the advantages of explicit memory mechanisms for long-term dependency capture across irregular visit intervals.

3. **Pretraining Importance**: GCN pretraining on single-visit diagnosis classification was critical, with performance dropping 30+ percentage points on CN to MCI detection without pretraining. This demonstrates the value of task-specific transfer learning.

4. **Network Structure Preservation**: Operating directly on functional connectivity matrices rather than hand-crafted features preserved rich spatial information about brain organization and allowed learning of which brain regions and connections matter most.

5. **No Imputation Needed**: The approach avoided missing data imputation by operating on complete functional connectivity matrices, simplifying the pipeline compared to approaches requiring handling of missing features.

6.2 Implications for Alzheimer's Disease Prediction

6.2.1 Model Architecture Matters

The thesis demonstrates that architectural choices substantially impact performance on this task. Specifically:

- **Attention-Based Processing**: Transformers' parallel self-attention mechanisms proved more effective than sequential RNN processing for identifying converter sequences, despite RNNs being specifically designed for sequential data.
- **Spatial-Temporal Decomposition**: The GNN-RNN combination effectively separated spatial brain network modeling from temporal sequence modeling, each handled by architectures optimized for their respective tasks.
- **Data-Modality Alignment**: Different architectures suit different data modalities. Tabular features work well with attention mechanisms; network-structured data benefits from graph neural networks.

6.2.2 Complementary Information Sources

The two approaches operate on fundamentally different information:

- **Clinical Features**: Cognitive tests measure function directly; MRI/PET measure structure and pathology. These have clear clinical relevance and are directly interpretable.

- **Functional Connectivity**: Reflects functional organization and network integration, capturing indirect markers of neurodegeneration. Less directly interpretable but potentially more sensitive to early network-level changes.

Both achieved competitive accuracy (82-83%), suggesting both contain substantial predictive information. Ensemble approaches combining both could potentially improve further.

6.2.3 Early Detection is Challenging but Possible

Both approaches achieved respectable performance on early detection:

- Transformer: Highest performance on shortest visit histories
- GNN: 68.8% accuracy on CN to MCI transition

However, these results also highlight that disease prediction remains genuinely difficult. Error rates of 20-30% even on optimized models reflect the inherent biological variability in disease progression and limitations of available data.

6.2.4 Recent Information is Most Valuable

Both approaches, through different analyses, converge on the finding that recent clinical information is more valuable than historical data:

- Transformer shows diminishing returns with longer histories
- GNN implicitly learns temporal importance through LSTM

This supports clinical intuition that the current disease trajectory is more predictive than distant history.

6.3 Methodological Contributions

Beyond the specific application to AD prediction, this work contributes methodological insights:

6.3.1 Handling Longitudinal Imbalance

The thesis demonstrated two complementary approaches to handling class imbalance in longitudinal prediction:

- **Dataset Balancing**: Removing samples from majority class, simpler but reduces training data
- **Focal Loss**: Reweighting examples during training, maintains all data but requires careful tuning

The trade-offs between these approaches merit further investigation.

6.3.2 Temporal Feature Encoding

Both approaches explicitly encoded temporal information:

- Transformer: Temporal features as input
- GNN: Explicit concatenation of visit intervals

This stands in contrast to approaches that implicitly encode time through position (RNNs) or position embeddings (standard Transformers). The explicit encoding's effectiveness suggests this approach warrants broader adoption.

6.3.3 Irregular Sequence Handling

Medical data naturally has irregular spacing. Both approaches successfully accommodated this through explicit temporal feature engineering rather than requiring preprocessing to enforce regular spacing.

6.4 Limitations and Considerations

6.4.1 Dataset and Generalization

1. **ADNI Cohort**: Both approaches use ADNI-derived data, potentially limiting generalization to other populations with different demographic characteristics or healthcare systems.

2. **Dataset Size**: The GNN approach is limited to 303 subjects with fMRI data, smaller than the Transformer's 1677. Larger fMRI datasets could strengthen conclusions.

3. **Validation Requirements**: External validation on independent cohorts is needed before clinical deployment.

6.4.2 Model Complexity and Clinical Feasibility

1. **Transformer Complexity**: The Transformer architecture, while powerful, is more complex to implement and deploy than simpler statistical models.

2. **GNN Requirements**: Requires fMRI data and preprocessing infrastructure (fMRIPrep), limiting applicability to sites without advanced neuroimaging capabilities.

3. **Interpretability Gaps**: While attention weights provide some interpretability, understanding why specific patterns lead to predictions remains challenging.

6.4.3 Clinical Practical Considerations

1. **False Positive/Negative Trade-offs**: Both approaches require clinical specification of acceptable error rates. The thesis does not determine optimal operating points.

2. **Cost-Benefit Analysis**: Clinical value depends not only on prediction accuracy but on available treatments and their costs. This thesis focuses on prediction without considering therapeutic options.

3. **Patient Heterogeneity**: Both approaches assume homogeneous disease processes. Subtype-specific models might improve performance.

6.5 Future Research Directions

6.5.1 Immediate Extensions

Ensemble Approaches

Given the complementary nature of the two approaches, several ensemble strategies warrant investigation:

1. **Output-Level Ensembles**: Combine predictions through learned weights, leveraging both models' strengths.

2. **Feature-Level Fusion**: Concatenate Transformer-derived and GNN-derived features into a single model.

3. **Multi-Task Learning**: Joint training on Transformer's multi-class task and GNN's binary conversion task with shared representations.

Temporal Extensions

1. **Multi-Step Forecasting**: Extend from next-visit prediction to multiple steps into the future (e.g., 6-month, 1-year, 2-year forecasts).

2. **Time-to-Event Prediction**: Estimate not just whether conversion occurs, but when (months or years until predicted transition).

3. **Trajectory Prediction**: Forecast the full cognitive trajectory, not just next state.

6.5.2 Longer-Term Directions

Multi-Modal Integration

Develop unified architectures that jointly process:

- Tabular clinical features
- Structural imaging (MRI volumes, cortical thickness)
- Functional imaging (fMRI connectivity)
- Molecular biomarkers (PET, CSF)
- Genetic information (APOE status, polygenic risk scores)

Modern architectures like Vision Transformers extended to multi-modal data could handle this complexity.

Uncertainty Quantification

Extend point predictions to include confidence intervals:

- Bayesian neural networks for posterior distributions
- Conformal prediction for distribution-free uncertainty
- Calibration analysis of model confidence

Clinical decision-making requires not only predictions but understanding prediction uncertainty.

Interpretability and Explanation

1. **Attention Analysis**: Deeper investigation of what Transformer attention learns about disease progression.
2. **GNN Node Importance**: Analysis of which brain regions and connections the model emphasizes.
3. **Feature Attribution**: Methods like SHAP or LIME to understand feature contributions to individual predictions.
4. **Counterfactual Analysis**: What changes in visits/biomarkers would change predictions?

Subtype Discovery

Rather than assuming homogeneous disease:

- Unsupervised learning to identify disease subtypes with different progression patterns
- Subtype-specific models with improved performance
- Understanding heterogeneity in disease mechanisms

6.5.3 Clinical Validation and Deployment

External Validation

1. Test on independent cohorts
2. Assess generalization across demographic groups
3. Evaluate performance in clinical settings, not just research datasets

Clinical Trial Integration

1. Prospective validation in ongoing clinical trials
2. Integration into clinical trial enrichment strategies
3. Assessment of whether model-identified high-risk individuals benefit from interventions

Software and Standards

1. Development of open-source implementations
2. Clear documentation and validation protocols
3. Standardization for reproducible clinical use

6.6 Broader Impact and Societal Implications

6.6.1 Clinical Impact

If validated in clinical settings, these approaches could:

- Enable earlier identification of individuals at highest conversion risk
- Support enrichment of clinical trials with high-risk populations
- Assist clinicians in communicating prognosis to patients and families
- Guide treatment intensification decisions

However, implementation requires careful consideration of:

- False positive implications (anxiety for stable individuals)
- Potential for algorithmic bias
- Equity in access to neuroimaging data

6.6.2 Algorithmic Fairness

The approaches should be evaluated for:

1. **Demographic Bias**: Do error rates differ by age, sex, education, ethnicity?
2. **Data Representation**: Are certain population groups underrepresented in training data?
3. **Feature Bias**: Do the features used encode or amplify existing healthcare disparities?

6.6.3 Privacy and Data Security

Deployment of neural network models raises:

- Privacy concerns about sensitive health data
- Security requirements for protected data storage
- Regulatory compliance (HIPAA, GDPR, etc.)
- Clear data governance policies

6.7 Final Remarks

This thesis demonstrates that deep learning approaches show genuine promise for predicting Alzheimer's disease progression. The comparison of Transformer and Graph Neural Network approaches reveals that no single best architecture exists; rather, architectural choices should align with data modality and specific prediction task.

The complementary nature of the approaches—one excelling at clinical feature integration, the other at brain network structure—suggests that future progress will come from developing unified approaches leveraging multiple data sources and architectural paradigms.

Beyond the specific AD prediction application, this work contributes methodological insights relevant to broader clinical prediction problems: how to handle irregular longitudinal data, incorporate temporal information, address class imbalance, and balance model complexity against clinical feasibility.

Ultimately, automated prediction can support but not replace clinical judgment. The error rates observed—even in optimized models—reflect genuine biological

complexity that no algorithm will fully capture. The goal is to augment clinician decision-making by identifying high-risk individuals likely to benefit from intensive monitoring and early intervention, not to replace clinical assessment. With continued development, validation, and responsible deployment, such approaches could contribute meaningfully to improving early detection and patient outcomes in Alzheimer's disease.

Bibliography

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bouzidi, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

CR Jack, MA Bernstein, NC Fox, P Thompson, G Alexander, and D Harvey. The alzheimer's disease neuroimaging initiative: baseline characterization. *Neurobiology of Aging*, 29(10):1185–1204, 2008.

Xi Li, Yifan Zhou, Nicha Dvornek, Muhan Zhang, Sheng Gao, Juntang Zhuang, Dustin Scheinost, R Todd Constable, and Lawrence H Staib. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74:102233, 2020.

Razvan V Marinescu, Neil P Oxtoby, and Esther E Bron. Tadpole challenge:

Prediction of longitudinal evolution in alzheimer's disease. *arXiv preprint arXiv:1810.03909*, 2018a.

Razvan V Marinescu, Neil P Oxtoby, Esther E Bron, et al. Tadpole challenge: Prediction of longitudinal evolution in alzheimer's disease. *arXiv preprint arXiv:1810.03909*, 2018b.

Mahdi Moghaddami, , Mohammad Siadat, Austin Toma, Connor Laming, and Huirong Fu. Predicting alzheimer's disease progression using rs-fmri and a history-aware graph neural network. In *Proceedings of SPIE-the International Society for Optical Engineering*, 25–27 Jun .

Mahdi Moghaddami, Clayton Schubring, and Mohammad Siadat. Transformer model for alzheimer's disease progression prediction using longitudinal visit sequences. In *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pages 804–816. PMLR, 25–27 Jun 2025. URL `https://proceedings.mlr.press/v287/moghaddami25a.html`.

Susanne G Mueller, Michael W Weiner, and Leon J Thal. Ways toward an early diagnosis in alzheimer's disease: The alzheimer's disease neuroimaging initiative (adni). *Alzheimer's and Dementia*, 1(1):55–66, 2005.

Philip Scheltens, Kaj Blennow, Monique MB Breteler, Bart de Strooper, Giovanni B Frisoni, Stephen Salloway, and Wiesje M Van der Flier. Alzheimer's disease. *The Lancet*, 388(10043):505–517, 2016.

The Lancet Editors. Alzheimer's disease: Early detection and intervention. *The Lancet*, 2020.

Ashish Vaswani, Noam Shazeer, Nik Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, pages 5998–6008, 2017.

Appendix A

Appendix

A.1 Detailed Statistical Analysis

A.1.1 Transformer Model Statistical Tests

Two-sample Welch's t-tests were performed to compare models across 10-fold cross-validation folds ($n_1 = n_2 = 10$). Key results:

Table A.1: Statistical significance of Transformer vs. RNN comparison (raw dataset)

Comparison	**Metric**	**p-value**
Transformer vs. GRU	Balanced Accuracy (overall)	0.041
Transformer vs. GRU	Sensitivity (overall)	0.039
Transformer vs. minRNN	Accuracy (overall)	0.015
TST vs. LSTM	Accuracy (converter)	1.48×10^{-10}
TST vs. LSTM	Sensitivity (converter)	1.13×10^{-6}

A.1.2 GNN Model Variation Across Folds

The high standard deviation in GNN balanced accuracy (11.4%) reflects fold-to-fold variability due to class imbalance. This variation is expected but highlights the importance of multiple cross-validation folds in robust evaluation.

A.2 Hyperparameter Details

A.2.1 Transformer Hyperparameters

- Embedding dimension: 128
- Number of attention heads: 8
- Number of Transformer blocks: 4
- Feed-forward hidden dimension: 512
- Dropout rate: 0.3
- Learning rate: 1e-4
- Batch size: 32
- Max epochs: 500
- Early stopping patience: 50 epochs

A.2.2 GNN Hyperparameters

- GraphSAGE hidden dimensions: [128, 64]
- TopK pooling ratios: [0.5, 0.3]

- LSTM hidden dimension: 128
- Dropout rate: 0.3
- Focal loss parameters: $\alpha = 0.9, \gamma = 3$
- Learning rate (tuned by Bayesian optimization)
- Batch size: 32

A.3 Data Processing Details

A.3.1 Transformer Data Preprocessing

Feature normalization applied per-fold to avoid data leakage:

1. Compute mean and standard deviation on training set
2. Apply same normalization to validation and test sets
3. MinMax scaling applied separately to temporal features

A.3.2 GNN Data Preprocessing

Functional connectivity matrix normalization:

1. Fisher z-transformation: $z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$
2. Per-subject normalization
3. Threshold applied: set very small correlations to zero for sparsity

A.4 Feature Distributions

Key feature statistics from the Transformer datasets:

Table A.2: Feature statistics for cognitive tests

Feature	Mean	Std Dev
MMSE	26.92	3.50
ADAS-Cog13	16.62	10.68
FAQ	4.65	6.96
MOCA	23.52	4.18

A.5 Ablation Study Details

A.5.1 Transformer Ablation Results

Table A.3: Impact of removing feature categories (balanced dataset)

Ablation	Balanced Accuracy	Change
Full model	0.839	—
Cognitive scores removed	0.706	-15.76%
MRI volumes removed	0.812	-3.21%
Biomarkers removed	0.832	-0.84%

A.6 Implementation Details

A.6.1 Transformer Implementation

Implemented using PyTorch Transformers library with custom temporal feature engineering. Key architectural decisions:

1. **Positional Encoding**: Standard sinusoidal encoding plus learned temporal features

2. **Sequence Handling**: Padding to max sequence length in each batch, attention masks applied

3. **Output Aggregation**: Global average pooling over sequence before classification

A.6.2 GNN Implementation

Implemented using PyTorch Geometric library with custom multi-visit handling:

1. **Graph Construction**: FC matrix thresholded at 0.3 correlation, normalized adjacency

2. **Batch Processing**: Time series of graphs processed as batches of graphs

3. **Memory Efficiency**: GraphSAGE sampling to reduce memory requirements for large graphs

A.7 Reproducibility

Code and preprocessed data are available at [repository URL will be provided]. To reproduce results:

1. Clone repository

2. Install dependencies: `pip install -r requirements.txt`

3. Download TADPOLE and ADNI data from appropriate sources

4. Run preprocessing scripts: `python preprocess_transformer.py`, `python preprocess_gnn.py`

5. Train models: `python train_transformer.py`, `python train_gnn.py`

6. Generate results: `python evaluate.py`

A.8 Software and Libraries

- PyTorch: Neural network implementation
- PyTorch Transformers: Transformer architectures
- PyTorch Geometric: Graph neural networks
- Scikit-learn: Machine learning utilities and metrics
- NumPy, Pandas: Data processing
- Matplotlib, Seaborn: Visualization
- fMRIPrep: fMRI preprocessing
- Nilearn: Neuroimaging utilities