Explainable Reinforcement Learning for Alzheimer's Disease Progression Prediction

Raja Farrukh Ali, Ayesha Farooq, John Woods, Emmanuel Adeniji, Vinny Sun, William Hsu

Kansas State University
{rfali, ayeshafarooq, jwoods03, adeniji, vinnysun1, bhsu}@ksu.edu

Abstract

We present a novel application of SHAP (SHapley Additive exPlanations) to explain the decisions made by Reinforcement Learning (RL) models used for Alzheimer's Disease (AD) progression prediction. Leveraging RL to predict the variation in brain cognition and model 10-year cognition trajectories using only baseline year-0 data, we employ SHAP to explain the model's decision-making process. Our approach provides detailed insights into the key factors influencing AD progression predictions, offering both global and individual, patient-level explainability. Our results show that while the RL model is able to achieve a mean absolute error on par with supervised learning methods, the model fails to properly capture established markers of Alzheimer's Disease such as amyloid accumulation when put through the lens of post-hoc explanation methods like SHAP. By bridging the gap between predictive power and transparency, our work empowers clinicians and researchers to gain a deeper understanding of AD progression and better understand machine learning methods proposed for healthcare, thereby facilitating more informed decision-making in AD-related research. Our code is available at https://github.com/rfali/xrlad.

Introduction

Alzheimer's disease (AD) is a progressive, irreversible, neurodegenerative disease that affects millions of individuals worldwide. AD is characterized by the progressive decrease in brain size and the eventual death of neurons. It causes memory deterioration, language deterioration, cognitive deficits, and impairments in judgment and communication. Understanding the factors driving AD progression and developing accurate prediction models are crucial for early diagnosis, intervention, and improved patient outcomes (Porsteinsson et al. 2021).

While machine learning models have shown promise in predicting AD progression, their lack of interpretability and explainability pose a significant challenge to their adoption (Vellido 2020). Interpretability refers to the ability to understand the internal mechanisms and workings of a model. A model is interpretable if humans can understand its predictions or decisions in a straightforward and intuitive manner. Interpretable models often have clear, transparent structures and make use of features that are easily understandable to humans. For example, linear models or decision trees are generally considered interpretable because their decision-making process can be easily traced and understood. Interpretability in supervised machine learning, where models are trained on labeled data, differs from unsupervised machine learning, where models learn patterns from unlabeled data. In supervised learning, interpretability focuses on understanding the relationship between input features and the model's predictions. However, in reinforcement learning (RL), interpretability becomes more challenging due to the inherent sequential decision-making nature of the problem, as well as the traditionally large state and action spaces associated with RL environments (Vouros 2022).

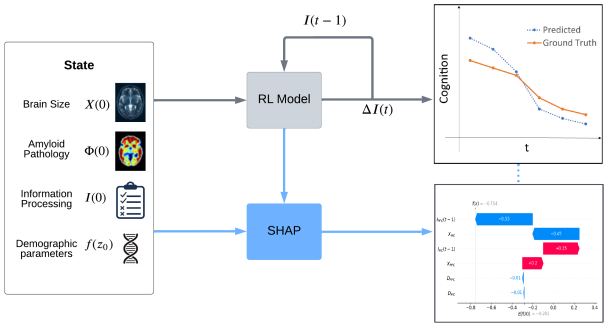


Figure 1: SHAP is used to explain an interpretable, domain knowledge based RL model's cognition trajectory predictions over 10-years for Alzheimer's Disease progression.

Explainability, on the other hand, refers to the ability to provide explanations for predictions or decisions made by a model. Even if a model is complex and not easily interpretable, it can still be considered explainable if it can provide explanations for why it made a particular prediction or decision. These explanations help users understand the factors or features that contributed to the model's output. Explainability techniques often involve generating post-hoc explanations, such as feature importance scores, attention maps, or textual explanations, to clarify the model's reasoning. Explainability in healthcare not only helps in understanding the factors that contribute to progression of a

disease but also enables clinicians to make informed decisions regarding patient care (Bogdanovic, Eftimov, and Simjanoska 2022).

In this work, we apply the SHAP (SHapley Additive exPlanations) method (Lundberg and Lee 2017) to enhance the explainability of an interpretable RL model proposed for AD progression prediction by Saboo et al. (2021). By modeling the brain as a system of differential equations based on causal relationships (hence interpretable), the RL agent learns to predict change in cognition (actions) as it tries to optimize cognition while minimizing associated costs. SHAP provides a method of assigning feature importance scores and quantifying the contribution of each feature to the RL decisions. By leveraging SHAP, we gain valuable insights into the relative importance of different features, shedding light on the factors driving AD progression in a more explainable manner. The main contributions of this work are:

- We propose an explainable, reinforcement learning based disease prediction model for Alzheimer's Disease that can predict 10-year cognition trajectories from baseline year-0 data.
- The proposed interpretable and explainable RL model offers both global and local level predictions and their associated explanations. The global level explanations identify and rank the most important featuresconsidered by the model when making cognition predictions over successive years on a dataset containing 1660 samples. The local explanations are per-sample, which can identify how the model predicts and explains its predictions for a particular patient's sample data.
- Among the three feature vectors, the model ranks the cognition score at the previous time step (year) as the most important feature in making the next prediction of the cognition score, followed by the brain region size, and lastly by amyloid accumulation in a region.
- The model's explainability aspect also highlights potential failure modes since the model, while being accurate in predicting long-term cognition trajectories, does not consider amyloid accumulation an important feature vector in its prediction, which is contrary to some clinical studies (Feld et al. 2014).

Related Work

Modeling Alzheimer's Disease Progression

Research into modeling the progression of Alzheimer's disease can be generally classified into mechanistic models and data-driven models. Mechanistic models rely on domain knowledge to encode relationships among variables through algebraic and/or differential equations (Vértes et al. 2012; Li et al. 2016; Frässle et al. 2018; Galioulline et al. 2023). Data-driven models encompass a spectrum of techniques, including Bayesian models (Fruehwirt et al. 2018), event-based models (Fonteijn et al. 2012), mixed-effects models (Oxtoby et al. 2018; Liu et al. 2023), and machine learning models (Lin et al. 2018; Tabarestani et al. 2018; Saboo et al. 2020). These models leverage biomarker data to establish connections between disease pathology, region size, cognitive function, and demographic factors. See Frizzell et al. (2022); Balakrishnan, Sreeja, and Panackal (2023) for recent literature reviews on the use of AI/ML for AD diagnosis. These techniques exhibit a relatively low reliance on domain-specific knowledge and are effective for short-term forecasting. More recently, Saboo et al. (2021) combined mechanistic models and reinforcement learning to propose a hybrid model that leverages RL's ability to model a sequential decision making problem and predict AD's progression over time based on baseline imaging/cognition data and demographic features.

Explainable RL (XRL)

Along with the increased application of RL to real-world problems, there has been a surge in research dedicated to explainability of these models (Puiutta and Veith 2020). XRL research can be categorized in various ways. Here we give a brief overview of the recent literature surveys and the various methods used to categorize XRL works.

The most prominent method of categorization splits XRL methods into (a) transparent and (b) post-hoc explainability methods (Arrieta et al. 2020). SHAP is categorized as a post-hoc explainability method, which uses interaction data to explain the model's predictions (Heuillet, Couthouis, and Díaz-Rodríguez 2021). SHAP uses the magnitude of influence from each variable in the environment after training to quantify the interactions between and contributions of each variable towards the final prediction of the model. As per the taxonomy by Qing et al. (2022), SHAP falls into the Model-Explaining and Explanation-Generating category. This category describes methods that generate explanations from the model without being explicitly self-explainable.

XRL methods can also be categorized into one of these categories: Feature Importance, Learning Process and MDP, or Policy-Level (Milani et al. 2022). In particular, the Feature Importance category is divided into (a) Learn Intrinsically Interpretable Policy, (b) Convert to Interpretable Format, and (c) Directly Generate Explanation. SHAP falls within the *Directly Generate Explanation* subcategory within *Feature Importance* category. SHAP generates an explanation after training from a non-interpretable policy. This enables the understanding of the factors that influence a model towards its final predictions.

Although SHAP is a very popular method owing to the widespread use of the associated library, it has seen very limited use in RL applications (Kumar, Vishal, and Ravi 2022; Raz et al. 2022), primarily because of the traditionally large state and action spaces associated with RL environments and perhaps also because of the lack of a standardized way for implementing RL models. Thus, we hope this work will be able to shed more light on the use of Post-Hoc Explanation-Generating tools such as SHAP in the growing field of XRL, especially in problem settings with low-dimensional state and action spaces.

Background

In Reinforcement Learning, an agent is tasked with the objective of sequentially interacting with a given environment

to accrue rewards over time, where actions leading to favorable outcomes are positively reinforced, while suboptimal actions incur penalties. The RL problem is typically formulated as a Markov Decision Process (Puterman 2014). The MDP is defined by a tuple $< S, A, P, R, \gamma >$, where S represents a set of states, A is the set of available actions, $P : S$ x $A \rightarrow \mathbb{P}(S)$ is the transition probability function, $R : S$ x $A \rightarrow \mathbb{R}$ is the reward function, and γ is the discount factor. By leveraging the Markov Decision Process (MDP), a formalized policy $\pi : S \rightarrow \mathbb{P}(A)$ can be established to describe the behavior of the agent. The policy predicts the probability of taking action a given state s by mapping states to a distribution of actions. From state s_t, which is sampled from the policy π, the agent receives a reward r_t and transitions to the next state s_{t+1}. In an episodic problem, the process continues until the agent reaches a terminal state. The expected total sum of discounted rewards by starting in state s and following policy π for the rest of the episode gives a value function $V^{\pi}(s)$. The goal of the agent is to find the optimal policy π^* that maximizes the expected discounted cumulative reward. This can be formulated as finding a policy π such that $V^{\pi^*}(s) \geq V^{\pi}(s)$ for all states s, where $V^{\pi}(s)$ is the expected discounted cumulative reward gained during an episode starting from state s while following policy π.

We now provide a brief overview of the RL model used to predict AD progression, based on the works of Saboo et al. (2021). The model leverages domain knowledge to establish causal relationships between various factors involved in AD progression. To summarize, Amyloid beta (Aβ), a key factor in AD and measured using florbetapir-PET, propagates between brain regions, influencing brain structure measured via MRI, activity measured via fMRI, and cognition measured through tests like Mini-Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale - Cognitive Subscale 11 and 13 (ADAS11 and ADAS13). The model defines a hypothetical variable, C_{task}, which represents cognitive demand and impacts brain activity. Brain activity, in turn, affects cognition and contributes to neurodegeneration. The model also considers the energetic cost associated with brain activity, which can further contribute to neurodegeneration.

The model defines these relationships using appropriate sets of differential equations (DEs). The brain is modeled as a graph $G_S = (V, E)$, where a node $v \in V$ represents a brain region, and an edge $e \in E$ represents a tract. Let $X_v(t)$ denote the size of a brain region $v \in V$ at time t, and $X(t) = [X_1(t), X_2(t), ..., X_{|V|}(t)]$. $D_v(t)$ is the instantaneous amyloid accumulation in region $v \in V$ at time t. The total amount of amyloid in a region is $\varphi_v(t)$. $Y_v(t)$ denotes the activity in region $v \in V$ in support of cognition $C(t)$ at time t. Although cognition, brain size (X_v), and activity (Y_v) are related, the exact relationship among them is unknown and cannot be easily learned from limited data. The energetic cost $M(t)$ represents the brain's energy consumption, which is proportional to its overall activity $Y_v(t)$ and serves as a cost associated with supporting cognition.

The brain is modeled as a graph $G_s = (V, E)$, where each node $v \in V$ represents a brain region, and each edge $e \in E$ represents a tract. Let $X_v(t)$ indicate the size of brain region $v \in V$ at time t, and let $X(t) = [X_1(t), X_2(t), \ldots, X_{|V|}(t)]$.

A network diffusion model is used to model change of amyloid in a region over time as it captures the propagation of Aβ through tracts. $D_v(t)$ is the instantaneous amyloid accumulation in region $v \in V$ at time t, the change in which can be represented as

$$\frac{dD_v(t)}{dt} = -\beta H D_v(t) \tag{1}$$

where $D(t) = [D_1(t), D_2(t), \ldots, D_{|V|}(t)]$, H is the Laplacian of the adjacency matrix of the graph G_s, and β is a constant. The total amyloid in a region $\phi_v(t)$ can then be expressed as

$$\phi_v(t) = \int_0^t D_v(s)\, ds \tag{2}$$

To support cognition, multiple brain regions work in synchrony. We denote the activity in region $v \in V$ in support of cognition $C(t)$ at time t as $Y_v(t)$. The hypothetical term information processing, $I_v(t) \in R_{\geq 0}$, is introduced to relate a region's size and activity to its "contribution" to cognition. The resulting model for cognition, $C(t)$, supported by the brain at time t can be modeled as

$$C(t) = \sum_{v \in V} I_v(t) \tag{3}$$

The activity, $Y_v(t)$, in a region depends on both its information processing and its size. The relationship between activity and information processing is proportional, while the relationship between activity and size is inversely proportional. The relationship between the three features is modeled as

$$Y_v(t) = \gamma \frac{I_v(t)}{X_v(t)} \quad \forall v \in V \tag{4}$$

The brain consumes energy to support cognition. The energy consumption for a region is proportional to the activity in that region. Therefore the total energy cost of the brain can be modeled as

$$M(t) = \sum_{v \in V} Y_v(t) \tag{5}$$

Neurodegeneration, the change in brain size, is influenced by two factors: amyloid deposition and brain activity. Previous equations and models inferred a linear relationship between the rate at which a brain region degenerates and Aβ deposition. Brain degeneration is also be accelerated by brain activity. The following equation is a representation of how brain activity, neurodegeneration, and Aβ are considered related in this model.

$$\frac{dX_v(t)}{dt} = -\alpha_v D_v(t) - \alpha_v Y_v(t) \quad \forall v \in V \tag{6}$$

The demographics of individual patients can affect the progression of Alzheimer's Disease. To account for the influence of demographics in the model, parameters α_1, α_2, β, γ were introduced in previous equations. For demographics

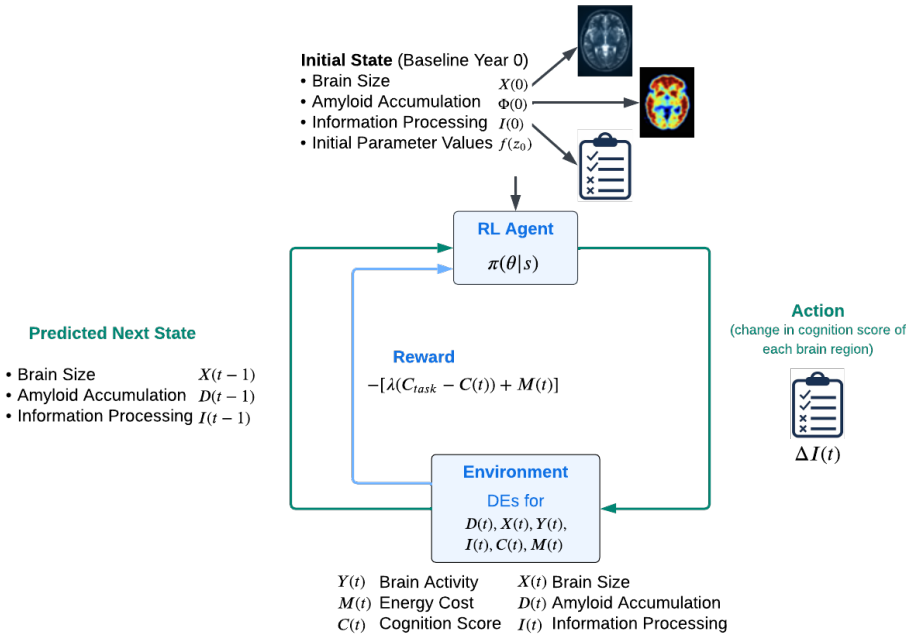


Figure 2: The RL Model. The state $S(t)$ includes the brain regions' size, amyloid accumulation, cognition (information processing) at the previous time step, and demographic features. The action $A(t)$ specifies the change in information processing for one time step (year). The environment simulates brain dynamics and captures causal relationships encoded using Differential Equations. The reward $R(t)$ balances cognitive demand and energy cost. During training, the agent trains on single time steps for 1M steps, while during evaluation, baseline year-0 data is used to generate 10-year cognition trajectories.

at baseline Z_0, let f be a function that approximates these parameter constants, such that

$$(\alpha_1, \alpha_2, \beta, \gamma) = f(Z_0) \tag{7}$$

We refer the reader to (Saboo et al. 2021) for more details on methods used for this parameter estimation.

DEs provide relationships between some, but not all, factors relevant to AD. To address missing relationships, the model formulates an optimization problem, which it solves using reinforcement learning. Figure 2 explains how the RL model works. The environment is represented as a simulator that encompasses the equations governing various factors, including $D(t)$, $\phi(t)$, $X(t)$, $Y(t)$, $I(t)$, $C(t)$, and $M(t)$. The state at time t, denoted as $S(t)$, comprises the current sizes of the brain regions $X(t)$, the Amyloid accumulation $D(t)$, and the information processed by each region at the previous time step $I(t-1)$. The action at time t, $A(t) \in \mathcal{A}$ specifies the change in information processed by each brain region from the previous time step, i.e., $\Delta I_v(t) \in \mathbb{R}\ \forall v \in V$. Formally, the state and action spaces are defined as:

$$S(t) = (X(t), D(t), I(t-1)) \tag{8}$$

$$A = (\Delta I_1(t), \ldots, \Delta I_{|V|}(t)) \tag{9}$$

$$I_v(t) = I_v(t-1) + \Delta I_v(t); \quad \sum_{v \in V} I_v(t) \leq C_{\text{task}} \tag{10}$$

The goal of the RL agent is to calculate the optimal information processing in each brain region, which all together add up to the total cognition of the brain. In order to do so, it must balance the trade-off between two competing criteria: (i) minimizing the discrepancy between the cognitive demand of a task C_{task} and the actual cognition available in the brain $C(t)$, and (ii) minimizing the cost $M(t)$ associated with supporting cognition. The reward $R(t)$ at time t is defined as follows, where λ is a parameter controlling the trade-off between the mismatch and the cost, and the agent's goal is to maximize this reward given by

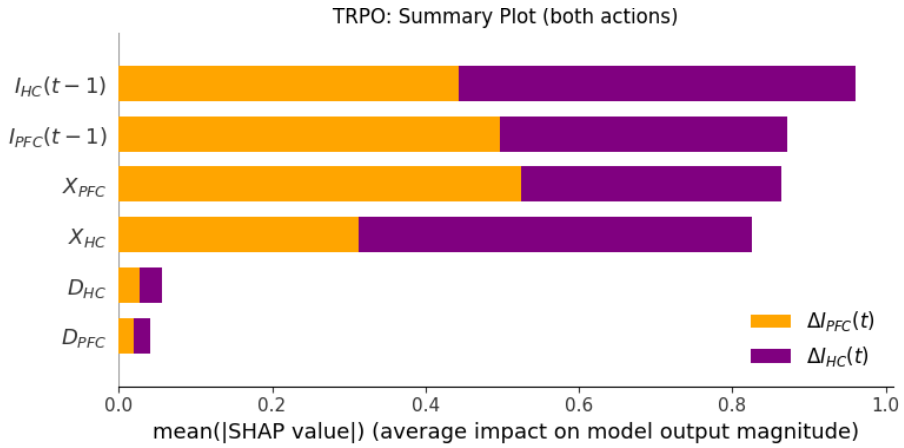
$$R(t) = -\left[\lambda(C_{\text{task}} - C(t)) + M(t)\right] \tag{11}$$

This is the objective function for the optimization problem of distributing cognitive workload optimally across all brain regions (Saboo et al. 2021). The model predicts disease progression by considering the interactions between the DEs, effectively creating a simulator, and the actions taken by the RL agent.

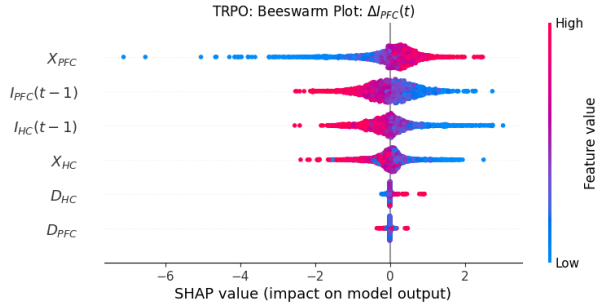
Using SHAP to explain RL model's predictions

Shapley Values

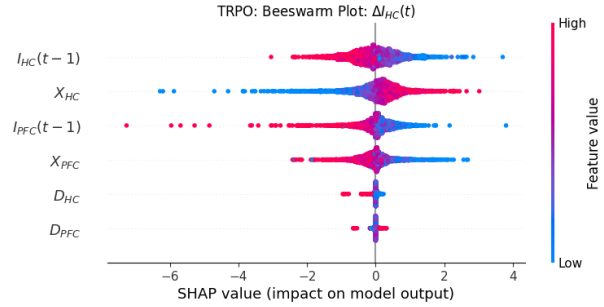
Shapley values (Shapley 1953), rooted in cooperative game theory, assign a value to each player based on their marginal contribution to different coalitions (subsets of players) to fairly allocate the total payoff of the game to each player. The Shapley value of a player is based on their marginal



(a) Summary Plot showing the ranked features by importance for the two model outputs



(b) Beeswarm plot for the action ΔI_{PFC}



(c) Beeswarm plot for the action ΔI_{HC}

Figure 3: SHAP Plots for all patient predictions highlighting the input features in the model's decision making. The model ranks the cognition measured at the previous time step as being the most important feature, followed by brain region size, and lastly amyloid accumulation. Moreover, (a) depicts the corresponding share of each feature for a given prediction class (shown in different colors). Figures (b) and (c) depict the beeswarm plots for each region, assigning distinctive colors to sample values (red high, blue low). Experiments were conducted for 1660 patient samples using 5-fold cross validation, and each fold was repeated 5 times with different random seeds.

contributions to all possible combinations in which they participate. Mathematically, the Shapley value for a player i in a game with N players is defined as

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{12}$$

where $\phi_i(v)$ is the Shapley value of player i, $v(S)$ is the value of the coalition, $|N|$ is the total number of players, and $|S|$ is the number of players in coalition S.

SHapley Additive exPlanations (SHAP)

Lundberg and Lee (2017) build on Shapley values and its extensions to develop a model-agnostic explainability framework called SHAP (SHapley Additive exPlanations), offering a robust and coherent approach to interpreting model predictions. In context of Eq 12, $\phi_i(v)$ is the Shapley value of a specific feature in the model, $v(S)$ is the prediction made by the RL model for a specific set of features, $|N|$ is the total number of the input features, and S ranges over all possible coalitions excluding feature i. This signifies that when calculating the Shapley value for a specific feature i, all possible combinations of features without i are considered. The value of S changes as different subsets of features are examined. This way, SHAP values attribute a model's prediction to distinct features, explaining their influence on the model's output.

Specifically, the SHAP framework can be used to achieve two types of explainability. *Global explainability*: By aggregating SHAP values computed for each individual instance across the entire dataset, the framework provides a comprehensive perspective on the behavior of the model in predicting AD across a diverse spectrum of cases. This can help to identify consistent features that significantly influence predictions. *Local explainability*: By delving into the process of individual predictions for AD and considering the unique impact of each input feature, the SHAP framework can provide microscopic insights into the rationale behind specific predictions. This is pivotal for understanding the model's decision-making on a case-by-case basis, shedding light on the prominent factors guiding predictions for individual pa-

tients. Figure 1 visualizes the context of our experiment, where we input the RL model as well as the state space $S(t)$ to the SHAP library and generate global and local explanations.

Experimental Setup

We used Alzheimer's Disease data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). We selected individuals with baseline measurements of cognition, demographics, MRI, and florbetapir PET scans, along with longitudinal cognitive measurements and at least 2 follow-up assessments comprising both PET and MRI scans, as recommended by Saboo et al. (2021). Follow-up visits were not required to be consecutive, and data spanning up to 10 years after baseline were retained. Cognitive assessments were preserved for all available time points up to and including year 10, irrespective of MRI/PET availability. This resulted in a dataset of 160 patients, encompassing 52 cognitively normal (CN), 23 with significant memory concern (SMC), 58 with early mild cognitive impairment (EMCI), and 27 with late mild cognitive impairment (LMCI). Demographic features included age, gender, education, and the presence of the APOE-ϵ4 genotype. Our analysis focused on a 2-node graph representation (GS) with nodes denoting the hippocampus (HC) and the prefrontal cortex (PFC) due to their relevance to cognition and Alzheimer's disease pathology. Hippocampal and prefrontal cortex volumes were used to represent brain structure ($X(t)$), with raw hippocampal volumes normalized and PFC volumes scaled by the median ratio of PFC to hippocampus. PET-scan derived Standardized Uptake Value Ratio (SUVR) values for PFC and hippocampus served as measures of Aβ deposition ($\phi(t)$). We utilized the Mini Mental State Examination (MMSE) score as a measure of cognition, adjusting it to range from 0 to 10. We used Trust Region Policy Optimization (TRPO) (Schulman et al. 2015) in this work (details in Supplementary).

Results

TRPO achieved a mean absolute error of 0.572, on par with baselines from supervised machine learning (Saboo et al. 2021). The primary aim of this section is to examine the SHAP plots, extracting information about the model's behavior, and analyzing if the model's behavior aligns with established research within the field of AD.

Global Explanations Global explanations are visualized using bar and beeswarm SHAP plots (Figure 3). The bar plot ranks features by mean absolute SHAP value for information processing in the Prefrontal Cortex region ($I_{PFC}(t = 0)$) and the Hippocampus region ($I_{HC}(t = 0)$). Features with higher mean absolute values are placed at the top, indicating their greater influence. The beeswarm plot assigns distinctive colors to sample values, illustrating how high and low feature values impact the model's behavior.

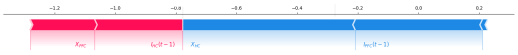
The bar plot in Figure 3(a) offers a global explanation of the RL model's predictions used in this work. The mean absolute value for each feature in each class is calculated and plotted. Saboo et al. (2021) attribute brain cognition to brain activity, $Y_v(t)$, and amyloid accumulation, $D_v(t)$. They also highlight the direct correlation between $Y_v(t)$, $X_v(t)$ and $I_v(t)$. Hence, with $I_v(t-1)$ and $X_v(t)$ being shown as the most important features, figure 3(a) illustrates brain activity having greater significance in the RL model's prediction of $\Delta I_v(t)$ than amyloid accumulation.

Similar to the bar plot, the beeswarm plots depicted in Figures 3(b) and 3(c) identify features $I_v(t-1)$ and $X_v(t)$ as being most impactful in the model's prediction of $\Delta I_v(t)$ for the two regions. Figure 3(b) shows that high feature values of X_{HC} and low feature values of $I_{PFC}(t-1)$, $I_{HC}(t-1)$, and X_{HC} increase the predicted $\Delta I_{PFC}(t)$. Figure 3(c) shows that low feature values for $I_{PFC}(t-1)$ and $I_{HC}(t-1)$ increase the model's prediction of $\Delta I_{HC}(t)$ while high feature values of X_{HC} increase the predicted cognition score. These plots help contextualize the model's behavior.

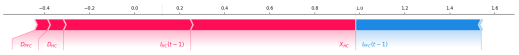
Local Explanations The SHAP plots used for local explanations include the decision, waterfall, and force plots (Figure 4). These plots display how input features affect the model behavior for a single sample. For the decision and waterfall plots, the expected value for all samples is plotted at the bottom. Each feature is then added to the plot starting with the least important feature at the bottom. Each feature has a SHAP value that pushes the expected value either positively or negatively. Once all features are added, the plot reaches the actual predicted value for that particular sample. This value is posted at the top of the plot. Features pushing the prediction higher are shown in red, whereas those pushing the prediction lower are in blue. The force plot serves the same function as the decision and waterfall plot but the plot is visualized along the x-axis.

Figure 4 explains how each feature impacts the model's prediction of $\Delta I_v(t)$ for the Prefrontal Cortex and the Hippocampus of a single patient at the baseline year. Figure 4(a), 4(c), and 4(e) explain the Prefrontal Cortex while Figure 4(b), 4(d) and 4(f) explain the Hippocampus region. The observed patterns in the global explanation plots are also visible in the plots for local explanations. We see that hippocampus size is the most important contributing factor to $\Delta I_{PFC}(t)$ in both brain regions, contributing positively in the Hippocampus and negatively in the Prefrontal Cortex. This is followed by the information processing of the Hippocampus at the previous timestep which contributes positively in both regions, and the information processing of the Prefrontal Cortex at the previous timestep, which contributes negatively. Amyloid accumulation in both regions has little effect. Interestingly, Prefrontal Cortex size has a significant effect on $\Delta I_{PFC}(t)$, but no effect on $\Delta I_{HC}(t)$.

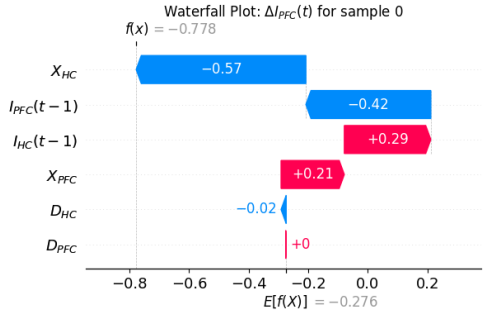
These global and local explanations point to the hypothesis that brain activity plays a more pivotal role in brain degradation than amyloid accumulation. Jagust and Mormino (2011) and Hampel et al. (2021) discuss the prominent Alzheimer's disease theory centered on β-amyloid (Aβ) protein deposition initiating cognitive decline. Jagust and Mormino (2011) explore the connection between lifelong brain activity patterns and Aβ deposition and suggest that manipulating neural activity could impact Aβ levels. This underscores the role of neural activity alongside Aβ in the cognitive decline process.



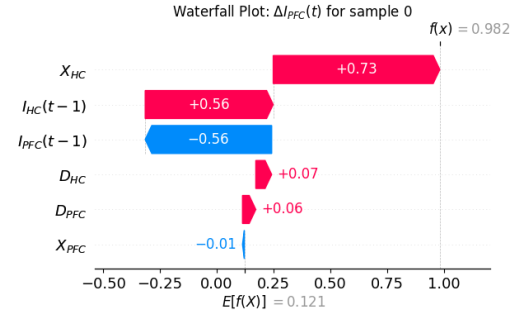
(a) Force Plot for $I_{\text{PFC}}(t=0)$



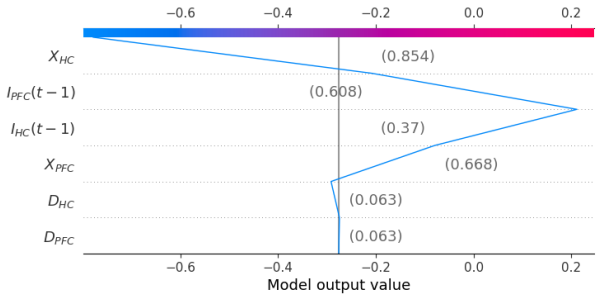
(b) Force Plot for $I_{\text{HC}}(t=0)$



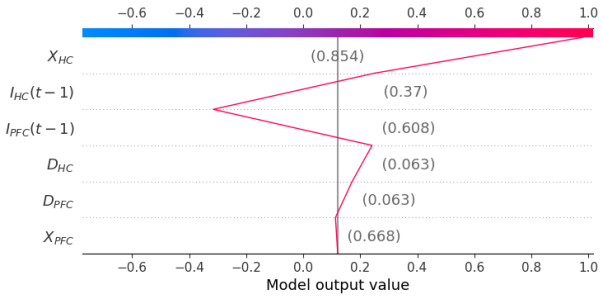
(c) Waterfall Plot for $I_{\text{PFC}}(t=0)$



(d) Waterfall Plot for $I_{\text{HC}}(t=0)$



(e) Decision Plot for $I_{\text{PFC}}(t=0)$



(f) Decision Plot for $I_{\text{HC}}(t=0)$

Figure 4: SHAP Plots for a single prediction: (a) and (b) Force Plots for Information processing $I_v(t=0)$ of both PFC and HC regions, (c) and (d) Waterfall Plots for both regions, (e) and (f) Decision Plots for both regions. Features pushing the prediction higher are shown in red, whereas those pushing the prediction lower are in blue.

Per-Patient Analysis

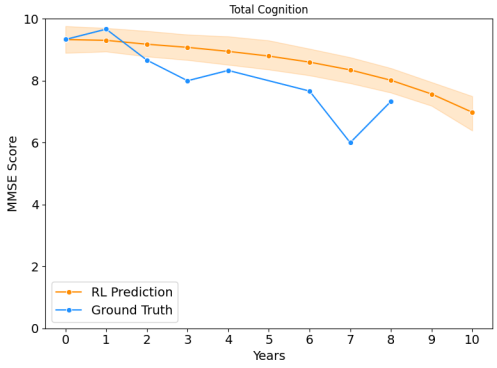
We also conduct an analysis on each patient to determine the effect these factors have on individuals as the disease progresses. Figure 5 shows the results for a particular patient (Patient Record ID 4294), who was selected for their maximum decrease in MMSE score. Figures 5(a), 5(c), and 5(e) show the accuracy of the RL model in predicting the total cognition $C(t)$, Prefrontal Cortex Size X_{PFC}, and Hippocampus Size X_{HC} respectively as Alzheimer's disease progresses in the patient.

Subfigure 5(b) shows how each feature affects the final SHAP value of the model, which corresponds to the change in cognition $\Delta C(t)$. In this patient, the information processing at the previous time step for Hippocmapus region has the greatest effect on the change in cognition, followed by the information processing of that region. It is also evident that the information processing of the hippocampus contributes more to the total cognition than the information processing of the prefrontal cortex in this patient.

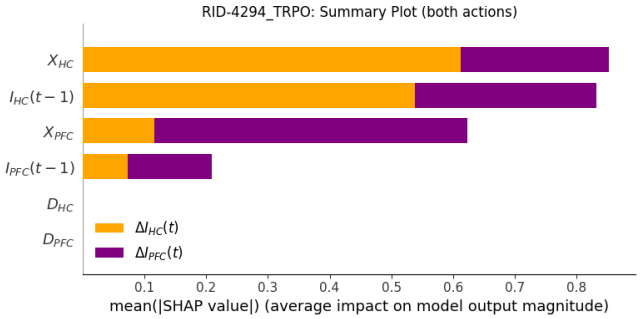
Figures 5(d) and 5(f) show the effect of X_{PFC} and X_{HC} in their respective regions on the change in cognition. For both regions, they initially contribute positively to the change in cognition, but as brain region size decreases, their effect also decreases and becomes negative. Guo et al. (2013) observed similar effects, where a larger initial brain size helped slow down Alzheimer's progression, but as degradation proceeded it also increased in rate. From these figures, we can conclude that Alzheimer's progression is dynamic and dependent upon the individual it affects.

Discussion

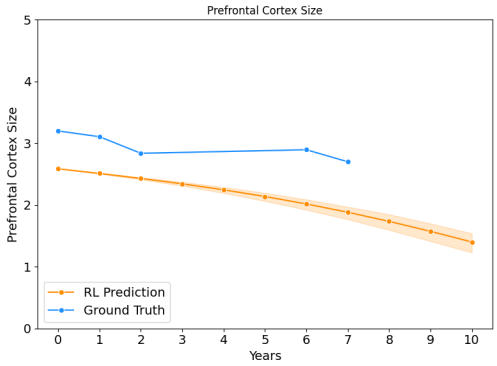
The swift progress of machine learning, notably with the rise of deep learning, has greatly influenced diverse domains. The integration of ML models in healthcare has been approached with caution, and for good reason. It is essential that these models are thoroughly understood and made explainable before they can be confidently utilized. We posit that the methodology presented in this work represents a step in this direction. By integrating machine learning models with explainability to predict cognition trajectories for early diagnosis, we aim to facilitate the gradual and manageable adoption of these tools. Limitation in capturing established



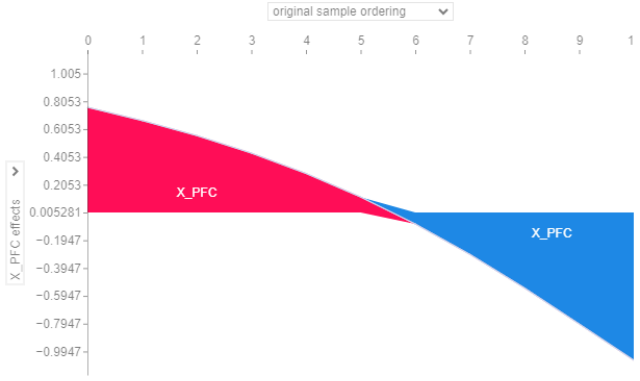
(a) RL Prediction vs. Ground Truth for Total Cognition



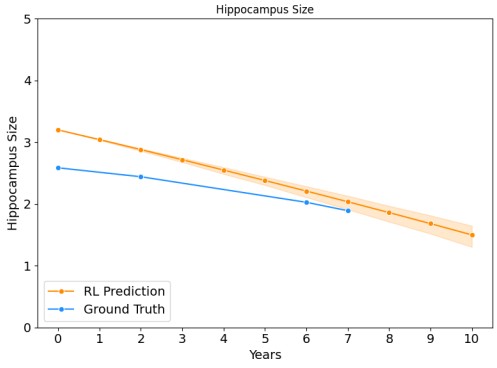
(b) Summary Bar Plot for $\Delta I_{PFC}(t)$ and $I_{HC}(t=0)$



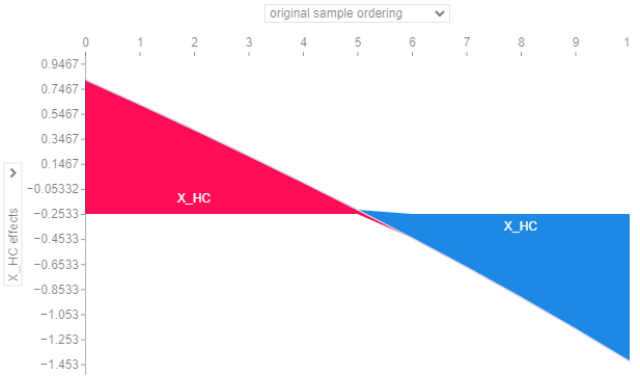
(c) RL Prediction vs. Ground Truth for PFC size



(d) Effect of PFC's size on Predicted PFC Cognition over 10 years



(e) RL Prediction vs. Ground Truth for HC size



(f) Effect of HC's size on Predicted HC Cognition over 10 years

Figure 5: RL Trajectories and SHAP Plots for a particular patient (Record ID 4294), aggregated across 5 seeds

AD markers like amyloid accumulation also highlights the need to improve these prediction models. Furthermore, the practical application of the insights gleaned from this work by clinicians could significantly enhance its value.

Conclusion

We demonstrate the use of RL and SHAP to predict and explain important features in Alzheimer's Disease progression. The reinforcement learning model employed can predict cognition trajectories up to 10 years post-diagnosis. SHAP analysis revealed that increased information processing and reduced brain region size significantly contribute to cognition decline in both of the studied brain regions. Our research aims to aid neurologists and researchers with Alzheimer's causality determination and treatment planning, especially early intervention and long-term prediction using clinical and demographic features. Future work can expand the scope by incorporating more brain regions to the model using domain knowledge. Moreover, extending this study on an even larger dataset would reveal its true performance and hidden biases. The ultimate goal is to translate these findings into actionable insights that can improve patient care.

Acknowledgments

This work was supported in part by the KDD Excellence Fund at Kansas State University. We gratefully acknowledge the anonymous reviewers for their critical and insightful feedback. Finally, this research was enriched by the broader academic community, and we appreciate the shared knowledge and resources.

References

Arrieta, A. B.; Diaz-Rodriguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lipez, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58: 82–115.

Balakrishnan, N. B.; Sreeja, P.; and Panackal, J. J. 2023. Alzheimers Disease Diagnosis using Machine Learning: A Review. *arXiv preprint arXiv:2304.09178*.

Bogdanovic, B.; Eftimov, T.; and Simjanoska, M. 2022. In-depth insights into Alzheimer's disease by using explainable machine learning approach. *Scientific Reports*, 12(1): 6508.

Feld, M.; Krawczyk, M. C.; Sol Fustinana, M.; Blake, M. G.; Baratti, C. M.; Romano, A.; and Boccia, M. M. 2014. Decrease of ERK/MAPK overactivation in prefrontal cortex reverses early memory deficit in a mouse model of Alzheimer's disease. *Journal of Alzheimer's Disease*, 40(1): 69–82.

Fonteijn, H. M.; Modat, M.; Clarkson, M. J.; Barnes, J.; Lehmann, M.; Hobbs, N. Z.; Scahill, R. I.; Tabrizi, S. J.; Ourselin, S.; Fox, N. C.; et al. 2012. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage*, 60(3): 1880–1889.

Frässle, S.; Lomakina, E. I.; Kasper, L.; Manjaly, Z. M.; Leff, A.; Pruessmann, K. P.; Buhmann, J. M.; and Stephan, K. E. 2018. A generative model of whole-brain effective connectivity. *Neuroimage*, 179: 505–529.

Frizzell, T. O.; Glashutter, M.; Liu, C. C.; Zeng, A.; Pan, D.; Hajra, S. G.; D'Arcy, R. C.; and Song, X. 2022. Artificial intelligence in brain MRI analysis of Alzheimer's disease over the past 12 years: A systematic review. *Ageing Research Reviews*, 77: 101614.

Fruehwirt, W.; Cobb, A. D.; Mairhofer, M.; Weydemann, L.; Garn, H.; Schmidt, R.; Benke, T.; Dal-Bianco, P.; Ransmayr, G.; Waser, M.; et al. 2018. Bayesian deep neural networks for low-cost neurophysiological markers of Alzheimer's disease severity. *arXiv preprint arXiv:1812.04994*.

Galioulline, H.; Frässle, S.; Harrison, S. J.; Pereira, I.; Heinzle, J.; and Stephan, K. E. 2023. Predicting future depressive episodes from resting-state fMRI with generative embedding. *NeuroImage*, 273: 119986.

Guo, L.-H.; Alexopoulos, P.; Wagenpfeil, S.; Kurz, A.; Perneczky, R.; Initiative, A. D. N.; et al. 2013. Brain size and the compensation of Alzheimer's disease symptoms: a longitudinal cohort study. *Alzheimer's & Dementia*, 9(5): 580–586.

Hampel, H.; Hardy, J.; Blennow, K.; Chen, C.; Perry, G.; Kim, S. H.; Villemagne, V. L.; Aisen, P.; Vendruscolo, M.; Iwatsubo, T.; et al. 2021. The amyloid-β pathway in Alzheimer's disease. *Molecular psychiatry*, 26(10): 5481–5503.

Heuillet, A.; Couthouis, F.; and Díaz-Rodríguez, N. 2021. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214: 106685.

Jagust, W. J.; and Mormino, E. C. 2011. Lifespan brain activity, β-amyloid, and Alzheimer's disease. *Trends in Cognitive Sciences*.

Kumar, S.; Vishal, M.; and Ravi, V. 2022. Explainable reinforcement learning on financial stock trading using shap. *arXiv preprint arXiv:2208.08790*.

Li, W.; Wang, M.; Zhu, W.; Qin, Y.; Huang, Y.; and Chen, X. 2016. Simulating the evolution of functional brain networks in alzheimer's disease: exploring disease dynamics from the perspective of global activity. *Scientific reports*, 6(1): 34156.

Lin, W.; Tong, T.; Gao, Q.; Guo, D.; Du, X.; Yang, Y.; Guo, G.; Xiao, M.; Du, M.; Qu, X.; et al. 2018. Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. *Frontiers in neuroscience*, 12: 777.

Liu, L.; Sun, S.; Kang, W.; Wu, S.; and Lin, L. 2023. A review of neuroimaging-based data-driven approach for Alzheimer's disease heterogeneity analysis. *Reviews in the Neurosciences*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Milani, S.; Topin, N.; Veloso, M.; and Fang, F. 2022. A survey of explainable reinforcement learning. *arXiv preprint arXiv:2202.08434*.

Oxtoby, N. P.; Young, A. L.; Cash, D. M.; Benzinger, T. L.; Fagan, A. M.; Morris, J. C.; Bateman, R. J.; Fox, N. C.; Schott, J. M.; and Alexander, D. C. 2018. Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain*, 141(5): 1529–1544.

Porsteinsson, A.; Isaacson, R.; Knox, S.; Sabbagh, M.; and Rubino, I. 2021. Diagnosis of early Alzheimer's disease: clinical practice in 2021. *The journal of prevention of Alzheimer's disease*, 8: 371–386.

Puiutta, E.; and Veith, E. M. 2020. Explainable reinforcement learning: A survey. In *International cross-domain conference for machine learning and knowledge extraction*, 77–95. Springer.

Puterman, M. L. 2014. *Markov decision processes*. Wiley Series in Probability and Statistics. Newy York: Wiley-Interscience.

Qing, Y.; Liu, S.; Song, J.; and Song, M. 2022. A survey on explainable reinforcement learning: Concepts, algorithms, challenges. *arXiv preprint arXiv:2211.06665*.

Raz, A. K.; Nolan, S. M.; Levin, W.; Mall, K.; Mia, A.; Mockus, L.; Ezra, K.; and Williams, K. 2022. Test and evaluation of reinforcement learning via robustness testing and explainable ai for high-speed aerospace vehicles. In *2022 IEEE Aerospace Conference (AERO)*, 1–14. IEEE.

Saboo, K.; Choudhary, A.; Cao, Y.; Worrell, G.; Jones, D.; and Iyer, R. 2021. Reinforcement learning based disease progression model for Alzheimer's disease. *Advances in Neural Information Processing Systems*, 34: 20903–20915.

Saboo, K.; Hu, C.; Varatharajah, Y.; Vemuri, P.; and Iyer, R. 2020. Predicting longitudinal cognitive scores using baseline imaging and clinical variables. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 1326–1330. IEEE.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International conference on machine learning*, 1889–1897. PMLR.

Shapley, L. S. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games*, volume II, 307–317. Princeton University Press.

Tabarestani, S.; Aghili, M.; Shojaie, M.; Freytes, C.; and Adjouadi, M. 2018. Profile-specific regression model for progression prediction of Alzheimer's disease using longitudinal data. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1353–1357. IEEE.

Vellido, A. 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24): 18069–18083.

Vértes, P. E.; Alexander-Bloch, A. F.; Gogtay, N.; Giedd, J. N.; Rapoport, J. L.; and Bullmore, E. T. 2012. Simple models of human brain functional networks. *Proceedings of the National Academy of Sciences*, 109(15): 5868–5873.

Vouros, G. A. 2022. Explainable deep reinforcement learning: state of the art and challenges. *ACM Computing Surveys*, 55(5): 1–39.

Supplementary Materials

RL Method

The environment of differential equations modeling the factors of Alzheimer's progression combined with the optimization of the objective/reward functions allows the use of model-free and on-policy RL methods to solve the optimization task. In this work, we used Trust Region Policy Optimization (TRPO) (Schulman et al. 2015), which is an on-policy algorithm that guarantees monotonic policy improvement. The key idea to TRPO is to constrain the local variation of the parameters to a "trust region" in the policy-space to ensure the update steps of the policy remains predictive. The constrain on the variation of parameter is determined by KL Divergence. The objective function can be described as:

$$\begin{aligned} max \qquad & E_{(s_t,a_t)\sim\pi}[\frac{\pi_\theta(a_t|s_t)}{\pi(a_t|s_t)}\hat{A}_\pi(s_t,a_t)] \\ \text{s.t.} \quad & D_{KL}(\pi_\theta(.|s)||\pi(.|s)) \leq \delta, \forall s \end{aligned} \tag{13}$$

TRPO is known for its stability and ability to handle complex, high-dimensional action spaces. In this work, TRPO acts as the learning agent that calculates the change in information processing for each of the two brain regions (the hippocampus and prefrontal cortex) for each time step (year) while optimizing the given reward function.

We provide the hyperparameters for our experiments in Table 1 and tabular results in Table 2.

Hyperparameters	**Values**
Batch size	1000
Epochs	1000
GAE λ	0.97
Max Cognition (C_{task})	10.0
Scale Observations	True
Action limit	$\pm$ 2.0
Score	MMSE
Max timesteps	11 (years)
Number of seeds	5

Table 1: Hyperparameters and configurations used in our experiments with TRPO

Result Metric	**Values**
Mean Absolute Error (test)	0.572
Mean Reward achieved (test)	-6.882

Table 2: Results achieved with TRPO