LEARNING AND COGNITION IN BRAIN AND MACHINE

Prediction of dementia from longitudinal data and modelling memory networks

Doctoral Dissertation by
Samaneh Abolpour Mofrad

Department of Computer Science,
Electrical Engineering and Mathematical Sciences

Faculty of Engineering and Science

Western Norway University of Applied Sciences

Spring 2021

Series of dissertations submitted to
the Faculty of Engineering and Science,
Western Norway University of Applied Sciences.

ISSN: 2535-8146
ISBN: 978-82-93677-64-2

Author: Samaneh Abolpour Mofrad
Title: Learning and Cognition in Brain and Machine

Cover:
Printed production: Molvik Grafisk AS, Bergen

Bergen, Norway, 2021

Dedicated with love to my parents,
Sedigheh Zarei & Mahmoud Abolpour,

who have shown me that it is never too late to pursue one's true passion.

PREFACE

The author of this thesis has been employed as a Ph.D. research fellow in the Engineering Computing research group at the Department of Computer science, Electrical engineering, and Mathematical sciences at Western Norway University of Applied Sciences, Norway. The research presented in this thesis has been conducted in cooperation with Mohn Medical Imaging and Visualization Centre (MMIV), Department of Radiology, Haukeland University Hospital, Bergen, Norway, and with the KTH Royal Institute of Technology, Stockholm, Sweden.

The author of this thesis has been enrolled the PhD program in Computer Science: Software Engineering, Sensor Networks and Engineering Computing, with a specialization in Engineering Computing.

This thesis is organized in two parts. Part I provides an overview of the relevant field and the background for the articles in the thesis, including a summary of the works. Part II consists of a collection of published and peer-reviewed research articles and submitted papers.

There are two collections of articles included in the thesis, of which the first is to be considered our main contribution.

Longitudinal data analysis and prediction of dementia:

Paper A Mofrad, Samaneh A., Lundervold, Arvid and Lundervold, Alexander S. 1 June 2021. A predictive framework based on brain volume trajectories enabling early detection of Alzheimer's disease. *Computerized Medical Imaging and Graphics*. Elsevier.

Paper B Mofrad, Samaneh A., Lundervold, Astri J., Vik, Alexandra and Lundervold, Alexander S. 22 January 2021. Cognitive and MRI trajectories for prediction of Alzheimer's disease. *Scientific Reports*. Nature Publishing Group.

Paper C Mofrad, Samaneh A., Bartsch, Hauke and Lundervold, Alexander S. From longitudinal measurements to image classification: Application to longitudinal MRI in Alzheimer's disease. *Under review.* April 2021.

Computational models for Memory networks:

Paper D Mofrad, Asieh A.*, Mofrad, Samaneh A.*, Yazidi, Anis and Parker, Matthew G. 6 April 2021. On neural associative memory structures: Storage and retrieval of sequences in a chain of tournaments. *Neural Computation*. MIT Press. *: equal contribution.

Paper E Mofrad, Asieh A., Yazidi, Anis, Mofrad, Samaneh A., Hammer, Hugo L. and Arntzen, Erik. 3 March 2021. Enhanced equivalence projective simulation: A framework for modeling formation of stimulus equivalence classes. *Neural Computation*. MIT Press.

ACKNOWLEDGMENTS

The time dedicated to researching and writing this dissertation has been my greatest academic adventure, and its completion required the encouragement, involvement, and support of many wonderful people. The following paragraphs attempt to express my deepest gratitude to those who were instrumental in bringing this project to fruition.

My greatest appreciation goes to my main supervisor, Assoc. prof. Alexander Selvikvåg Lundervold, for all of his support and trust. I am honored to have had the opportunity to work with Alexander, a kind and smart supervisor. I appreciate his guidance and insight, especially his eagerness to discuss new ideas, thoughts, and methodologies. Additionally, my project would not have been connected to the Mohn Medical Imaging and Visualization (MMIV) Center without his help. This connection has provided me with opportunities to meet many interesting people with different backgrounds and knowledge. Beyond Alexander's great and invaluable academic advice, he has become a trustworthy friend with whom I can freely share my thoughts and concerns.

My appreciation also extends to my co-supervisor, Prof. Yngve Lamo. I am grateful for his kind support, his contribution of time and ideas during the final stages of my work, and his exceptional comments when finalizing my dissertation. I would also like to express my gratitude to Olivier Verdier, my co-supervisor during the early stages of my PhD, who gave me the opportunity for a research stay in the mathematics group of KTH university in Winter 2017 and Spring 2018. Furthermore, the "advanced statistics in deep learning" course that I took at KTH proved significant in the formulation of my PhD.

An immense thank you goes to Prof. Arvid Lundervold and Prof. Astri J. Lundervold who introduced me to the remarkable research area of brain aging and cognition. I have learned so much from their great expertise and experiences in this field, and I am honored for the opportunity to collaborate with them in publishing papers together.

I am particularly thankful to the head of the Department of Computer Science, Electrical engineering, and Mathematical sciences, Assoc. prof. Kristin Fanebust Hetland. She is steadfastly supportive and solution-oriented, and kindly available for solving the problems at work, as well as personal concerns of us as PhD candidates. I feel fortunate that such a kind, powerful, and supportive woman was the head of our department during my work. Furthermore, I am thankful to Prof. Håvard Helstrup, the coordinator of the PhD program, who has always been available to resolve any problems and concerns throughout these four years. His consultations have always been supportive and have facilitated this accomplishment and the completion of my doctoral degree.

Working at MMIV center was a great pleasure for me. I am grateful to the colleagues at MMIV center, especially to the gracious, competent, and, humble leader of the center, Prof. Dr. Med. Ingfrid Haldorsen, for creating a friendly and enjoyable environment. I would like to extend my gratitude to Hauke Bartsch who always has new ideas to discuss and generously shares his knowledge with people around him in a friendly and patient manner. Collaborating with him in Paper C was an excellent opportunity for me, and I learned much from this collaboration, especially in the area of statistics. I also thank Marek Kocinski for his kind availability in helping me. I should also convey my plants' gratitude to him, as he has watered and taken care of them whenever I was absent.

I would like to take this opportunity to express my appreciation of Heather Arghandeh Paudler, my clever and kind-hearted friend, for her generous help during the last two years, especially for reading and commenting on part of my work.

During my PhD when I didn't have the chance to visit my wonderful family, my great friends have always supported me and have made this time more pleasant and memorable. I would especially love to mention Zahra Khorsand, Mahdi Roozbahane, and their lovely 5-year-old son, Rastin, who have always been like a little family for me. With Zahra I feel that I have an older, supportive, and trustworthy sister, and with Rastin, who is a cute little friend for me, I have always felt that I have access to a source of energy and happiness here in Bergen.

Throughout the years, I have been uplifted and supported by my family members. Above all, I am grateful to my parents. No words can express my heartfelt appreciation for all the love, unfaltering encouragement, and meaning they have given me. They have been my advocates and consultants, and I believe that I am where I am today because of them. Some years ago when I was in a difficult situation, I began a masters program at UiB in part to make them happy. From them I have learned that happiness comes from helping others; when nothing makes you happy, try to make someone else happy. This concept can change one's perspective, and through it I have found happiness. My parents' encouragement has inspired me to take risks and follow many dreams, including this one.

My extended family has been a constant source of strength, peace, and happiness throughout this journey of life and continue to inspire me. A special thanks goes to my supportive brothers Sajad and Khosrow, my sister Asieh, my sister-in-law Ghazal, and my brother-in-law Iman. I would like to express my gratitude to my older brother Sajad, who has always been my go-to person for technical questions especially about computer issues, for his patience and persistence in resolving them.

Finally, last but not least, my deepest thanks go to my sister Asieh for always being my best friend, confidant, the first one I seek for consulting, and the most available when I need someone. I am magnificently lucky that she lives in Norway, and I am actually here because of her. I would like to sincerely thank her for involving me in Paper E, her impressive, intriguing work, which allowed me to confront a new field of research. My research has been enriched by hers, including our experience in writing Paper D, our joint efforts during July 2018 and continuing within Fall 2020, and our 8-10 morning Zoom meetings all were an excellent experiment which was only possible with her special support. Working with such a kind, smart, accurate, and motivated person is a great pleasure for me.

Finally, I gratefully acknowledge the support and financing from the Western Norway University of Sciences and the Mohn Medical Imaging and Visualization Center.

Samaneh Abolpour Mofrad
Bergen, May 2021

ABSTRACT

Starting in the mid-20th century and throughout their developments, modern neuroscience and artificial intelligence (AI) have provided each other with inspiration, insights, and tools. The degree to which they are intertwined has been in constant flux over the years, but always present. With the enormous resurgence of interest in machine learning over the past decade, led by the much-celebrated successes of artificial neural networks and deep learning, the bond between the two fields seems to be growing stronger.

Artificial intelligence and machine learning have always kept an eye on *biological* intelligence and learning, as these provide our only examples of general intelligence and strong learning capabilities, inspiring the development of their much less capable–albeit improving–counterparts, which are based on computational models. The growing attention to both neuroscience and AI is also leading to growth where they intersect, i.e. in *neuroscience-inspired AI* and *AI-inspired neuroscience*, and in the usage of computational AI models within neuroscience and the cognitive sciences.

In this context, the present thesis aims to make a modest contribution through our application of machine learning techniques to the study of dementia using data from longitudinal MRI and psychometric testing, and through our proposed models for simulating aspects of the formation of memory networks during learning and memory retrieval. The former is our main contribution and is addressed in studies A, B, and C, while the latter is reflected in studies D and E.

Through longitudinal studies, i.e. studies based on the collection of repeated measurements from the same subjects, or experimental units, over time one can observe how measurements develop and discover new relationships between variables. Longitudinal data analysis is a large field of research comprised of a multitude of methods and is widely applicable to e.g. behavioural analysis and medicine. One inherently longitudinal phenomenon of particular interest for the present work is the biological, neurological, and cognitive alteration linked to aging. There is an immense need to develop methods that can indicate the risk of developing aging-related diseases such as dementia, as well as for increasing the understanding that is derived from new computational models for cognitive skills such as memory and learning.

The first part of this thesis (studies A, B, and C) develops and evaluates methods for using machine learning models with longitudinal data that have a time-dependent structure. We propose two novel and flexible frameworks to describe the trajectories of change extracted from the longitudinal data. The two frameworks are, respectively, based on (i) a combination of mixed-effects models in order to extract features from the longitudinal trajectories that can be used to train any type of machine learning classifier and (ii) mapping the multi-dimensional data onto two-dimensional images, enabling classifications based on convolutional neural networks.

The second part of this thesis (studies D and E) aims to construct simple and flexible models that can be used to simulate learning and memory retrieval processes in the human brain. These proposed memory networks are: (i) defining a new associative memory for storing sequences and investigate how to make efficient retrievals, and (ii)

a combination of a reinforcement learning model to form memory connections in the training phase and an iterative diffusion process to update the memory network to be used in the test phase.

We found that the frameworks proposed in the first part of the thesis, although being relatively simple approaches to the complexities of longitudinal data analysis, are comparable to other approaches in the literature as regards accurately predicting dementia. The proposed model for learning and retrieval based on associative memory in Paper D has several features that make it resemble its biological brain counterpart more than comparable models in the literature do, while significantly reducing errors in sequence-retrieval. The model for episodic memory developed in Paper E is quite flexible and can provide simulations of actual experiments on typical and atypical human behaviours.

Thesis at a Glance

Learning and Cognition in Brain and Machine

Prediction of dementia from longitudinal data and modelling memory networks

Part I

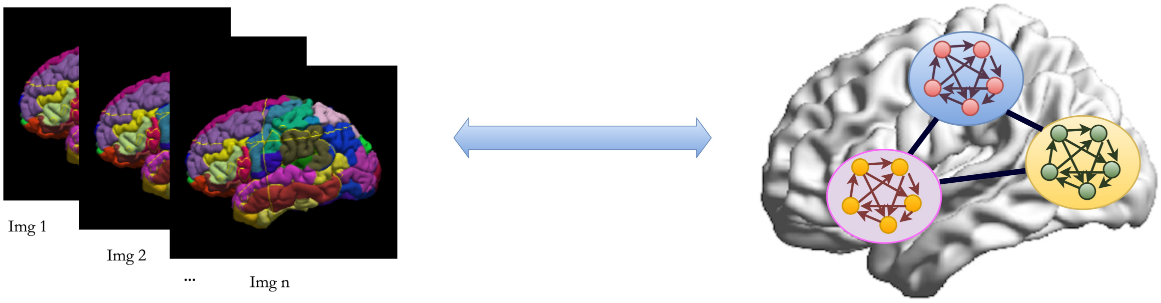
Constructing models for predicting the risk of dementia based on longitudinal data

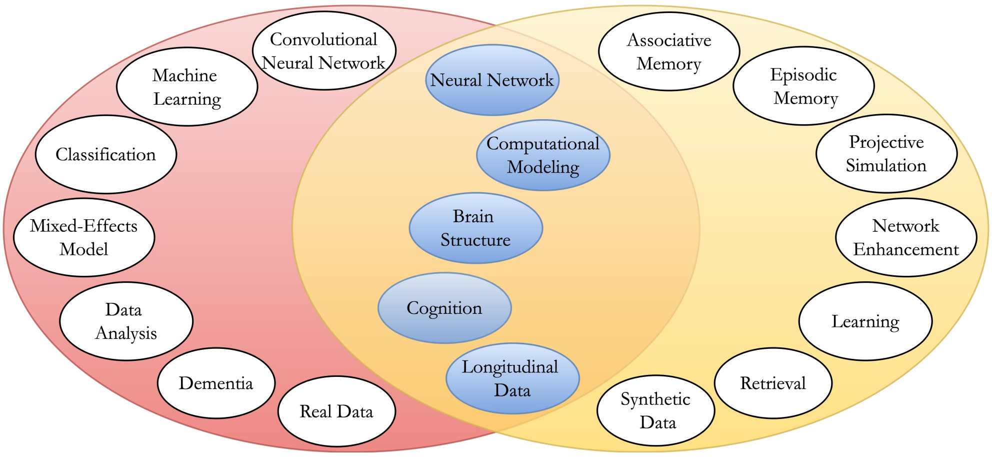
Part II

Modeling the process of learning and retrieval in episodic memory

Brain Aging

Computational Modelling





Convolutional Neural Network

Machine Learning

Classification

Mixed-Effects Model

Data Analysis

Dementia

Real Data

Neural Network

Computational Modeling

Brain Structure

Cognition

Longitudinal Data

Associative Memory

Episodic Memory

Projective Simulation

Network Enhancement

Learning

Retrieval

Synthetic Data

The main keywords for the two research directions of this thesis are presented in the Venn diagram, where some common concepts in the two parts are pictured in blue.

SAMMENDRAG

I løpet av deres utvikling siden midten av det 20. århundret har moderne nevrovitenskap og kunstig intelligens (AI) utvekslet inspirasjon, innsikter og verktøy. Graden av sammenfletting av de to feltene har variert gjennom årene, men den har alltid vært til stede. På grunn av den enorme interessen for maskinlæring over de siste ti årene, utløst og ledet av de velkjente suksesser innen kunstige nevrale nettverk og deep learning, ser båndet mellom de to feltene ut til å stadig bli sterkere.

Kunstig intelligens og maskinlæring har alltid holdt et blikk på *biologisk* intelligens og læring, da disse er våre eneste eksempler på generell intelligens og gode evner til læring, noe som inspirerer utvikling av deres mye mindre kapable analoger basert på beregnings-modeller. Den økende interessen for både nevrovitenskap og AI har også ledet til økt interesse for skjæringen mellom disse, i.e. for *nevrovitenskap-inspirert AI* og *AI-inspirert nevrovitenskap*, og også økt bruk av beregninsorienterte AI-modeller innen nevrovitenskap og kognitive vitenskaper.

Denne avhandlingen forsøker å gi et lite bidrag i denne kontekst, gjennom vår bruk av maskinlæringsteknikker innen studiet av demens basert på data fra longitudinell MRI og psykometrisk testing, og gjennom våre foreslåtte modeller for å simulere aspekt ved dannelsen av hukommelses-nettverk under læring og minneinnhenting. Det første temaet er vårt hovedbidrag, adressert i arbeidene A, B og C, mens det andre reflekteres i arbeidene D og E.

Gjennom longitudinelle studier, det vil si studier der en samler inn repeterte målinger fra samme individ, eller eksperimentell enhet, over tid, så kan man observere hvordan målingene utvikler seg og oppdage nye sammehenger mellom variabler. Longitudinell dataanalyse er et stort felt bestående av en lang rekke metoder, med et bredt anvendelsesområde innen for eksempel studiet av atferd og innnen medisin. Et iboende longitudnelt fenomen som er spesielt aktuelt for vårt arbeid er biologiske, nevrologiske og kognitive endringer forbundet med aldring. Det er stort behov for metoder som kan brukes til å indikere risiko for utvikling av demens og andre aldersrelaterte sykdommer, og også for en økt forståelse avledet fra nye beregningsorienterte modeller for ulike kognitive egenskaper slik som hukommelse og læring.

Avhandlingens første del (arbeidene A, B og C) utvikler og evaluerer teknikker for å bruke maskinlæringsmodeller sammen med longitudinelle, tidsavhengige data. Vi foreslår to nye og fleksible rammeverk for å karakterisere trajektorier, eller baner, avledet fra longitudinelle data. De to rammeverkene er (i) en kombinasjon av mixed effects-modeller for å trekke ut egenskaper fra longitudinelle baner som så kan brukes til å konstruere maskinlæringsbaserte klassifikatorer, og (ii) en representasjon av multi-dimensjonale data som to-dimensjonale bilder for å muliggjøre bruk av standard bildeklassifikasjonsmodeller slik som todimensjonale konvolusjonelle nevrale nettverk.

Avhandlingens andre del (arbeidene E og D) forsøker å konstruere enkle og fleksible modeller for å simulere lærings- og hukommelsesinnhentings-prosesser hos mennensker. Våre foreslåtte hukommelses-nettverk er basert på (i) en ny form for assosiativ hukommelse-struktur som kan lagre og effektivt finne igjen sekvenser, og (ii) en kombinasjon av en forsterkende læring-modell for å danne hukommelses-

forbindelser under en trenings-fase og en iterativ diffusjonsprosess for å oppdatere hukommelses-nettverket under en test-fase.

Vi fant at rammeverkene utviklet i avhandlingens første del gav prediktive modeller for demens med et lignende ytelsesnivå som andre modeller fra litteraturen, på tross av deres relativt enkle tilnærming til kompleksiteten i longitudinell dataanalyse. Modellen for læring og gjenfinning basert på assosiativ hukommelse utviklet i arbeid D har flere egenskaper som gjør den mer sammenlignbar med aspekter ved dens biologiske analoger enn andre sammenlignbare modeller fra litteraturen, mens den samtidig gir en betydelig reduksjon i antall feil under sekvens-gjenfinning. Modellen for episodisk hukommelse utviklet i arbeid E er relativt fleksibel, og kan gi simuleringer av faktiske eksperimentelle studier av typisk og atypisk atferd.

Contents

Part I

OVERVIEW

"In fact, the belief that neurophysiology is even relevant to the functioning of the mind is just hypothesis. Who knows if we're looking at the right aspects of the brain at all. Maybe there are other aspects of the brain that nobody has even dreamt of looking at yet."

—Noam Chomsky [22]

CHAPTER 1

INTRODUCTION

1.1 Motivation

Artificial intelligence (AI) and modern neuroscience have been growing together since the last century. Inextricably linked, with many early pioneers who worked simultaneously in AI and in neuroscience or psychology, they inspire, support, and give each other a boost. While AI and neuroscience have collaborated less in more recent years, on account of the growth in each field, they always keep an eye on one another, aspiring to grow at their intersection as well [60, 126].

Artificial Neural Network (ANN) is an approach to artificial intelligence that is used in deep learning, and is based on a (vastly) simplified model of biological neural networks [60, 84]. ANN architectures replicate the hierarchical structure of mammalian cortical systems with an information flow in consecutive and nested processing layers [82]. The neurosciences, in parallel with mathematical and logic-based methods, have always been a prolific source of inspiration for developing new algorithms and architectures for AI. For example, human learning through the senses of vision, hearing, and touch, have inspired the designs of ANN architectures and methods to train them by adjusting parameters via interactions and learning procedures in order to minimise error or maximise reward. In general, when a computational model replicates a cognitive behaviour, the model is a reasonable candidate for use in AI systems [60, 139].

Neuroscience attempts to explore how the brain system works and explain a wide variety of perceptual, visual, cognitive, and logical tasks [126]. The classical framework for neuroscience systems models simple computations in a neural system. This oversimplification is limiting and necessitates new approaches such as using the advantages of networks designed to learn from data. In ANNs, the computations are performed by specifying the structure of the network architecture and setting some learning rules, instead of designing a specific computational model [126]. In general, AI that enables computers to solve complex cognitive tasks can assist in the development of theoretical and experimental progress in neuroscience. The success of ANNs has encouraged cognitive scientists to use ANNs to investigate the biological cognition and its neural basis [24, 126]. Many recent findings have shown that AI models improve theories about the brain [60, 126]. For instance, they help to shed light on many behavioural and neuropsychological phenomena [77], replicates the transformation in perceptual systems [75], improves the structure of learning models that mimic memory functions [132], and helps to analyse and explain large amounts of neurobiological data [126].

When exploring cognitive sciences, a good model needs to be both explanatory and interpretative. An important property of ANN models is their predictive power, which helps achieve practical aims such as following changes related to disease in medical applications. For example, deep ANNs can be used to predict the brain action of a patient who has a damaged brain region as a result of neurodegenerative disease. Predictive power is an important component towards obtaining successful models in neuroscience [24].

It is within this context, that the present thesis seeks to make a modest contribution to the use of AI. We study cognitive behaviour and the prediction of future states of brain functionality, and design computational models for such phenomena. The main contribution of this work is the development of predictive models based on machine learning and deep learning for detecting the risk of dementia. Longitudinal data retains the dynamics of the variables in a study, which results in a greater predictive powers as it is possible to monitor the progression of the variables. In the second part of our work, we propose two models for episodic memory networks to simulate and analyse the learning and retrieval processes in memory, both brain and computer. A short summary of parts I and II are provided in the following sections.

1.2 Summary of Part I

The first part of this thesis (studies A, B, and C) deals with the challenges when using standard machine learning and deep learning methods on longitudinal data. These challenges include issues related to strongly inter-correlated variables, the presence of missing observations at different time-points, and the fact that subgroups of patients are often unbalanced. See Section 3.1 for definitions of the different types of longitudinal data.

To address these challenges, the thesis proposes two analytic frameworks that describe the trajectories of individual changes in longitudinal data. Descriptive features are extracted from the trajectories, and are then used to train machine learning and deep learning classifiers. In this part of the thesis, the frameworks are developed in order to identify features that predict mild cognitive impairment (MCI), and Alzheimer's disease (AD) at an early stage. Participants and measurements were taken from the relatively large and comprehensive ADNI dataset (`adni.loni.usc.edu`); the measurements included psychometric test scores and data from Magnetic Resonance Imaging (MRI) examinations. Figure 1.1 illustrates the two proposed frameworks and their application in the prediction of dementia.

The first framework (Fig 1.1: **a**) was designed as a combination of mixed-effects models, a class of models capable of producing regression models from dependent variables, which were used to derive features from MRI examinations and some cognitive tests, and machine learning models, which were used to predict dementia. The idea behind the second framework (Fig 1.1: **b**) was to convert the classification problems from the longitudinal data into image classification problems. This framework mapped the extracted trajectories of multi-dimensional data onto two-dimensional images, and then used these to train a convolutional neural network to perform image classification.

While in both approaches we used the trajectories of changes in the brain, in the first model, we extracted features from the trajectories and applied classical classification machine learning models, and in the second framework, we used the entire trajectories by mapping their values to an image and applied newer deep learning methods. With the data that we have used in these studies, both frameworks performed comparably.

The results indicate that the proposed approaches to longitudinal data analysis were successful in classifying the different subgroups of participants recruited in the ADNI study.

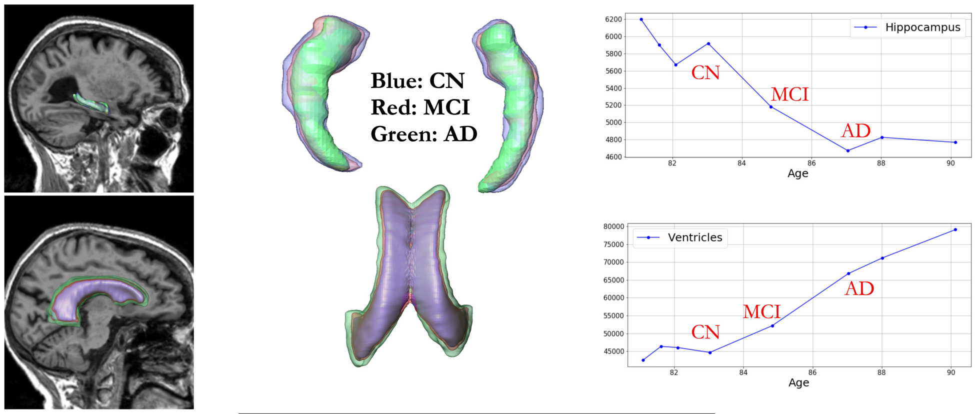
1.3 Summary of Part II

In the second part of this thesis (studies D and E), two flexible and interpretable computational models for learning and retrieval in associative- and episodic memories are proposed (Fig. 1.2). The model in paper D provides a new structure for associative memory, with feedback in the learning structure of a tournament-based neural network [72] (see section 2.2.2 for details). This allows retrieving the whole sequence from a given segment of it, regardless of where the segment is located in the sequence. Moreover, the newly proposed retrieval methods increase the model's efficiency and reduce errors when retrieving sequences, and make them more biologically plausible compared with earlier tournament-based neural networks [72] (Fig. 1.2: Paper D). The model in Paper E has two parts: a type of reinforcement learning model (projective simulation) with an episodic memory that simulates the learning procedure, and then an iterative diffusion process that updates the episodic memory so that indirect relations are derived, which is similar to the formation of equivalence relations in the brain (Fig. 1.2: Paper E). This model is quite flexible and can replicate the actual experiments on typical behaviour and the non-formation of relations in atypical groups such as among autistic children.

While the proposed models in Part II are simplifications of human memory, they could improve understanding and explaining of some types of cognitive behaviour.

1.4 Thesis organisation

This thesis has two components: Overview and Articles. The overview consist of six chapters, including an introduction (Chapter 1) that provides an overall picture of the thesis. Chapter 2 provides general insights into the relevant concepts in neuroscience, neural networks, and neuroimaging. Chapter 3 briefly explains longitudinal data, mixed-effects models, machine learning, and convolutional neural network as the basis of the work in Part I. Chapter 4 provides an overview of the data and software libraries used in this work. A summary of the five papers is given in Chapter 5. Finally, Chapter 6 draws some conclusions and suggests directions for future research. The papers are included in Part II.



(Paper A & B)

SID	IID	ROI1	ROI2	…	CogT1	CogT2	…	Age	Female	Education
1	Img1	1245	3481	…	21	37	…	68	0	8
1	Img2	1392	3443	…	-	33	…	69	0	8
1	Img3	1264	3529	…	22	37	…	71	0	8
1	Img4	894	2753	…	-	44	…	72	1	8
1	-	-	-	…	28	45	…	73	1	8
2	Img5	1026	3086	…	16	38	…	89	1	4
3	Img6	972	3006	…	18	28	…	74	1	15
3	Img7	1049	2988	…	-	32	…	75	1	15
3	Img8	918	3047	…	22	33	…	76	1	15
4	Img9	1010	2917	…	30	48	…	65	1	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1460	Img7421	932	2755	…	51	48	…	80	0	20

(Paper C)

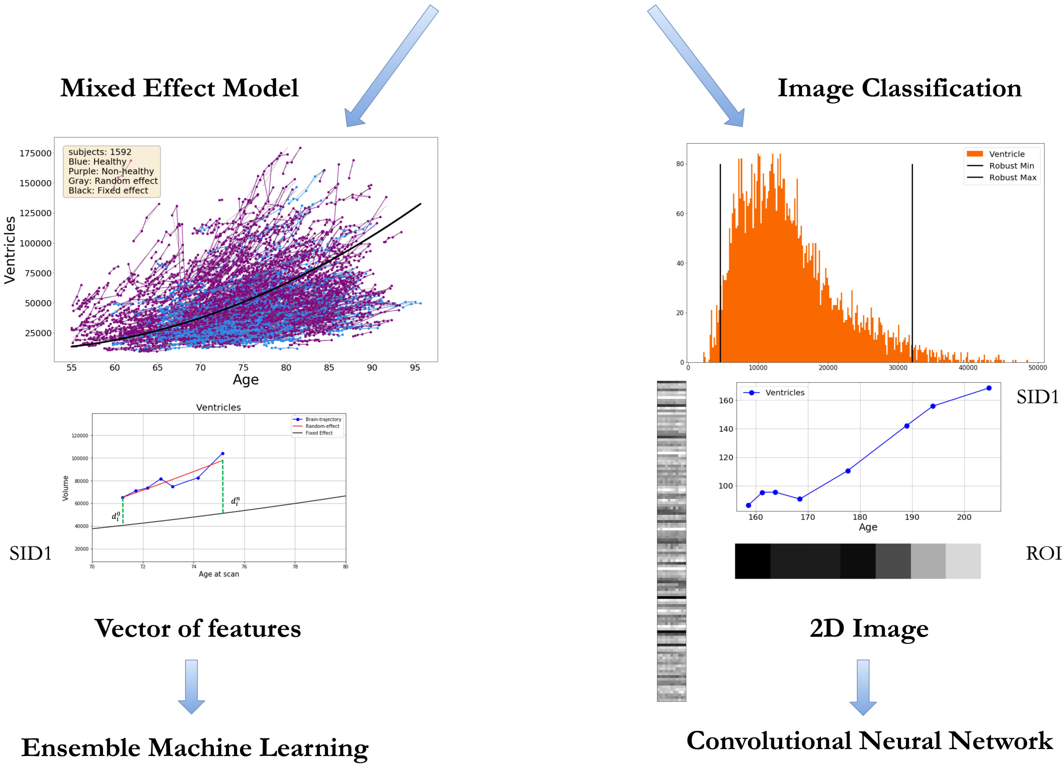


Fig. 1.1: Part I. The longitudinal data presented in this part of the thesis measured the development of brain volume and cognition during aging with MRI and cognitive tests, respectively. The volumes of the hippocampus and ventricles of one participant, who was cognitively normal (CN) before the age 85, and then suffered MCI, and at the age of 87 developed AD, are illustrated on the top. There is atrophy from aging and developing dementia in the hippocampus volume, whereas the ventricles are growing. An example of data is presented in the table in which SID and ROI is short for subject ID and regions of interest, respectively. The two frameworks in Part I have been designed to extract these alterations in the brain, represented in the table, to predict MCI and AD by using the mixed-effects model in Papers A and B, and by mapping the data onto 2D images in Paper C before applying machine learning and deep learning methods.

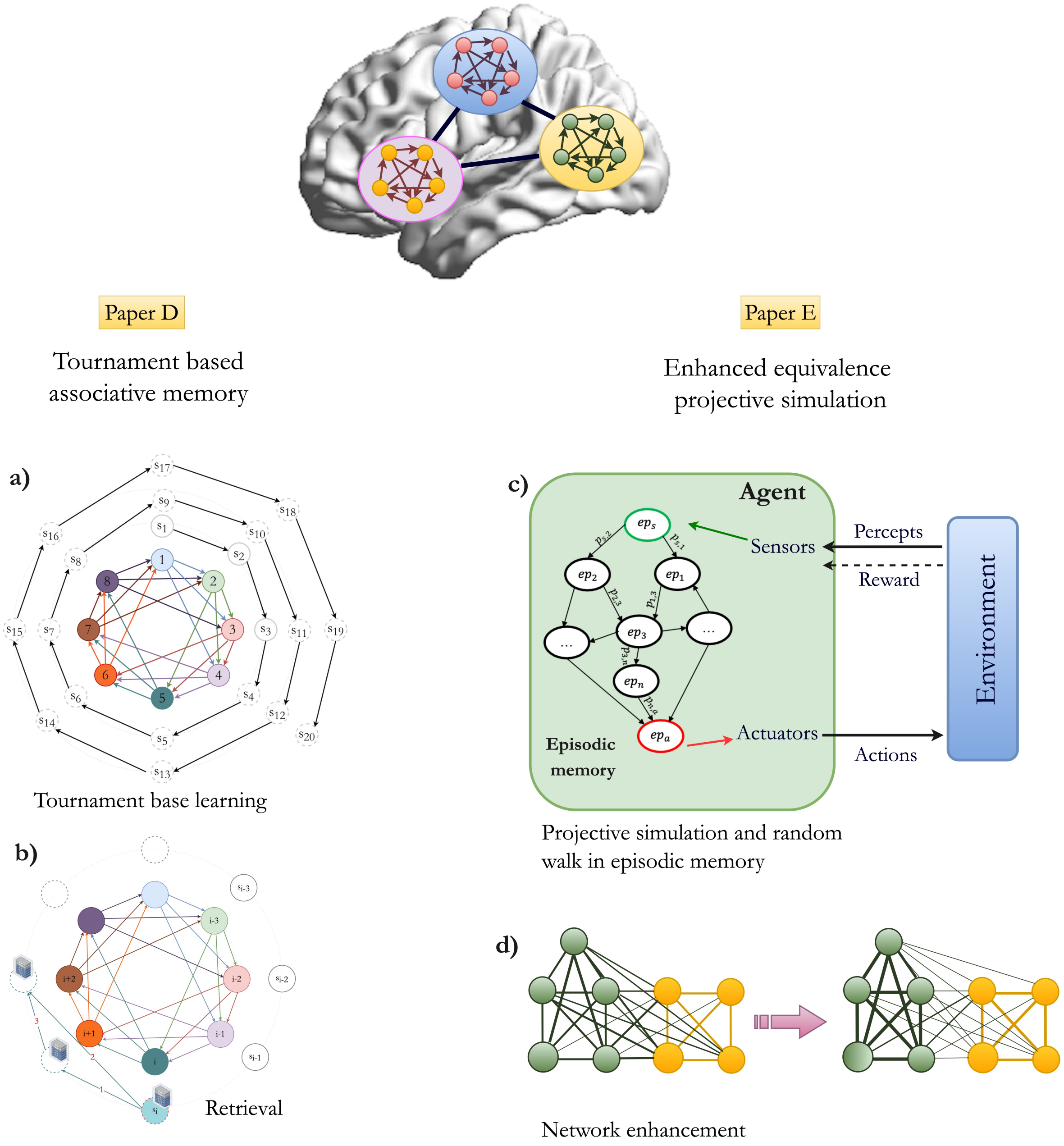


Fig. 1.2: Part II. Some computational models for memory and simulating the process of learning and retrieval are represented in Part II. In Paper D, a model for associative memory has been designed to store a sequence of information (here with length 20, $s_1 - s_{20}$) as a chain of tournaments (a). This paper suggests different algorithms for retrieving a sequence, given a segment of that sequence (see Section 2.2.2 for more details) (b). In Paper E, the projective simulation, that is a type of reinforcement learning containing episodic memory is used for the learning step (c); then a network enhancement procedure is employed to update the network of the episodic memory that formed (d) (see Sections 2.2.3 and 2.2.4 for more details). The figures presented for Paper D are from [99] and the figures for Paper E are based on figures presented in [15].

"Modules of brain networks define communities of structurally and functionally related areas, but they do not represent or support discrete mental faculties."

—Olaf Sporns [144]

CHAPTER 2

NEUROSCIENCE AND NEURAL NETWORKS

This chapter consists of two sections describing the essential terms in this thesis related to neuroscience and neural networks. Section 2.1 relates to memory, the learning process, aging, and neurodegenerative disease, and presents some neuroimaging techniques and psychometric tests. Section 2.2 presents definitions and tools for modelling networks such as artificial neural networks, associative memory, projective simulation, and network enhancement.

2.1 Brain structure and function

Brain connectivity has been studied by way of a combination of brain mapping techniques, including morphometry, diffusion tensor imaging and task-based or resting state functional MRI [131]. In this section we give a short presentation on the process of memory and learning in the brain, and its relevance to neurodegenerative diseases and dementia. This is followed by a presentation of some neuroimaging tools. Lastly, we present the two cognitive tests included in study B and their relations to the brain atrophy [107].

2.1.1 *Memory and learning*

Memory function is a process by which information is encoded, stored, and retrieved. There are many subtypes and models of memory function, and it is common to make a distinction between sensory memory, short-term memory, working memory, and long-term memory [6, 135]. In this model, sensory memory is important for sensory experiences obtained over very brief periods of time; short-term memory is important for retaining information over a short period of time (from a few seconds to a few days); working memory [7, 139] represents the ability to manipulate information temporarily stored in short-term memory; and long-term memory is important for preserving memory for a long time. Together, these partly overlapping aspects of memory function are of key importance to our functioning in daily life. This is also true of forgetting, which is of crucial importance to the replacement and updating of knowledge [56, 139].

Memory traces are not stored in specific areas of the brain, but the hippocampus is known as a sort of gating mechanism for information between the short- and long-term memory functions [14, 57, 128]. It is, however, important to take into account that there are multiple sensory and cognitive functions, as well as functional brain networks, involved when solving a memory task. Information must first reach and be interpreted by our sensory system, and several other cognitive domains are involved before the information is stored and recalled [139]. All these different aspects of memory function may be affected in a patient suffering from a neurodegenerative disorder; e.g., short-term memory is affected at an early stage of Alzheimer's disease. Modelling these memory functions and dysfunctions has been one of the most active subjects for researchers modelling brain networks, both for clinical and for theoretical use [99, 142].

Episodic memory, a part of long-term memory, is a major neurocognitive memory system that enables one to remember and elaborate ideas such as the self, subjective time, and consciousness. It is a past-oriented memory system that makes possible a sort of mental time travel from the present to the past and thus allowing one to remember and re-experience one's own previous experiences [149]. Associative memory is an episodically based memory system that forms the links in episodic memory [30, 38]. It refers to learning and remembering the relationship between initially unrelated items, such as an area and its name. Associative memory is heavily relied on in daily life [93, 143], but it is also one of the first parts of memory that is impacted by aging and neurodegenerative diseases such as dementia [93]. Associative memory has been an active topic for constructing memory models inspired by the human brain, aiming to both understand brain function and building memory stores for technical use [99, 142].

2.1.2 Neurodegeneration and dementia

Worldwide, average life expectancy has improved substantially over the previous decades [130]. In 2018, for the first time in recorded history, the number of people older than sixty-five surpassed the number of children five years or younger. Currently, about one in eleven people in the world are more than 65 years old. This is expected to grow to one in six by 2050. Furthermore, the number of people older than 80 is predicted to rise from 143 million in 2019 to 426 million in 2050 [151].

Aging is linked to a decline in processing speed, working memory, and inhibitory function, as well as atrophy in several brain structures [16, 113]. These normally-appearing changes are intensified in individuals with age-related diseases such as dementia [125], and make the discrimination between normal and disease-related aging both challenging and important [125].

About 50 million people worldwide suffer from dementia [161], with over 9.9 million new cases of dementia being diagnosed each year [121]. In 2016, dementia was the fifth leading cause of death in the world [111], and the number of people with dementia is predicted to be about 131 million by 2050 [121]. Since 2018, the overall cost of dementia care has grown to more than one trillion US dollars [121, 161], and dementia has a strong impact on public health, family members, and caregivers. Based on a study in the US [4], in 2018, more than 16 million family members and other unpaid caregivers were engaged for an estimated 18.5 billion hours taking care of people with dementia. Furthermore, this caregiving, which is valued at approximately $234 billion in the US, increases the risk of emotional distress, negative mental outcomes, and even physical suffering [4].

In 1906, Alois Alzheimer, a clinical psychiatrist, and neuroanatomist reported what he referred to as a severe and peculiar disease process affecting the cerebral cortex and behaviour of a 50 year old woman [64], a disease which is now called Alzheimer's disease (AD). AD is a common irreversible neurodegenerative disorder [5, 113], responsible for up to 60% to 70% of all cases of dementia [121]. Being a chronic neurodegenerative disease, AD causes a progressive death of brain neurons. As none of the available pharmacologic treatments and non-pharmacologic therapies can slow or stop the damage and destruction of neurons, even for moderate AD, treatment and therapies for preventing damage in the first place is crucial to slowing the disease process [4, 34, 105]. This early stage of AD lies partly within the construct of mild cognitive impairment (MCI) [117, 118].

MCI is characterised by a specific pattern of cognitive decline and represents a transitional state between normal cognitive aging and dementia [46]. An amnestic pattern of MCI is associated with an up to ten-fold increased in risk of AD [13, 116], but it is important to emphasise that not all patients with MCI will develop this disease. The search for biomarkers and other strong predictors of conversion from MCI to AD is, therefore, an important field of research [117, 162].

2.1.3 Brain structures and MR imaging

Brain imaging is one of the main tools for analysing and understanding the brain from a structural and functional point of view. For the purposes of the present study, we are most concerned with magnetic resonance imaging (MRI). This is a powerful technique for high-resolution 3D imaging of the human body, including the brain. MRI uses nuclear magnetic resonance to capture the anatomic structures and physiological function of tissues from a microscopic and molecular level [23, 110]. It is non-invasive, operating without harmful X-ray, and ionising radiation [26, 110]. MRI is one of the most broadly used techniques for monitoring and diagnosing the risk and process of diseases in the body [26].

So-called T1-weighted images, constructed based on a specific type of MR pulse sequences, are often used for the evaluation of anatomic structures, such as brain morphometry [23, 148]. In studies A, B, and C, we used T1-weighted images to monitor changes in brain volume in order to analyse the risk of dementia. Several studies have shown that different areas in the brain, such as the hippocampus and entorhinal cortex are associated with a disturbance of episodic memory and are affected by aging and dementia [20, 85, 123, 127]. These findings are in line with studies that indicate that the hippocampus and the entorhinal cortex are important brain regions of the brain for learning and which are also the first part of the brain damaged by neurodegenerative disease [93, 128].

Functional and diffusion magnetic resonance imaging are types of MRI-based imaging that take two different approaches to capturing the connectivity and motion in the brain [68, 73]. Functional MRI (fMRI) is an activity measurement that depends on blood oxygenation levels in the vessels and demonstrates regional metabolism changes in the human brain. The techniques of fMRI allows the mapping of brain function and has been widely applied in neuroscience research, including the monitoring of clinical and pharmacological interventions in cognitive diseases [48, 92]. Diffusion MRI is determined by detecting and measuring the diffusive motion of water's molecules in fluid-containing structures [18, 41]. These modern neuroimaging techniques help us investigate the connectivity and processing in the brain that result from cognitive behaviours [141, 152].

2.1.4 Cognitive tests

Cognitive tests are designed to measure more or less specific aspects of cognitive function [58], and performance on these tests can provide a tool to quantify aspects of brain function [78]. In Study B, we included the Rey Auditory Verbal Learning Test (RAVLT) [133] and the Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog) [129]. These tests are known to address some of the impairments in patients with AD and can be related to brain structures [107].

RAVLT is an episodic memory test assessing immediate and long-term memory function [39, 78, 133] The test is brief, easy to understand, and straightforward, and therefore it is acceptable for people aged 7 to 89 years [133]. This test is widely used as an early predictor of an amnestic type of MCI, and several studies have shown that impairment in RAVLT scores is reflected in atrophy measures obtained by neuroimaging examinations [107, 160].

In this test, the examiner reads a list of 15 unconnected words out loud at a rate of one word per second. The participant is then asked to recall as many words as possible. After five repetitions of this trial, the examiner reads a new list of 15 words, and the participants have to repeat as many words of the new words as they can remember. Immediately following this, the individual is asked to recollect as many words as possible from the initial list (trial 6). After a duration of 30 minutes, the participants are again asked to remember the words from the first list (trial 7) [107, 133]. For an elderly participant, the test can be rather stressful, revealing impaired cognitive abilities [28].

ADAS-Cog aims to assess the level of cognitive and non-cognitive behavioural symptoms associated with Alzheimer's disease through a short battery of tests [129]. It includes both subject-based tests and observer-based assessments. ADAS-Cog estimates cognitive domains such as memory function, language, and praxis through tasks including word recall, word recognition, naming objects, reasoning, and comprehension of spoken language [83]. In Study B, ADAS-Cog is used as a global measure of cognitive function. Higher scores on ADAS-Cog score means more severe impairment of cognitive function. The results of this test are expected to be linked to global neuroimaging markers of dementia [104, 107].

2.2 Computational models

There is immense interest in constructing models for brain networks, as they can provide insights into brain functionality, its memory systems, learning process and more. Additionally, such models can result in new techniques for computer science and artificial intelligence [11, 15, 101]. Even simple models can be useful in this respect, both as tools for exploring and assessing brain function, and especially as computational models able to solve relatively complex tasks.

2.2.1 Artificial neural networks

Artificial Neural Network (ANN) is a computational model that combines statistical techniques and graph networks, aiming to imitate biological neural networks. With the enormous increase in computational power and the amount of data generated across society, it has become a very powerful way to solve a variety of hard problems [47, 81, 96, 156].

The basic structural units of artificial neural networks are neurons and their connections, through which information is transferred. Brain neural networks process information in parallel through hundreds of billions of interconnected neurons; however, ANNs through an immense simplification of biological neural networks, nevertheless demonstrate a fairly good level of what can arguably be called intelligence, which can be sufficient for them to provide a better understanding of processes in the brain [60, 81, 156].

Learning, or signal propagation in ANNs, is a kind of automated weight change that operates locally and uses only the information of a few connected neurons to compute new weights. Furthermore, neurons operate in parallel and relatively independently of each other [81]. The locality and parallelism features make ANNs biologically plausible and allow them to execute much more quickly [81]. ANN models are characterised by their learning process, their connection patterns, and the weights associated with the connections, since these factors characterises the information flow through the network [153].

ANNs are well developed with a mathematical and statistical theoretical foundation, including linear algebra and Bayesian statistics [108, 109]. When a signal is transported through a connection, it is multiplied by the weight of the corresponding edge, which can be positive or negative [80, 153]. Some networks model more complex behaviour by connecting the final output nodes with earlier nodes, which causes the network to have feedback resulting in a highly nonlinear function, which is to a certain extent analogous to the behaviour of a brain neural network [96]. The setting of weights in a network is essential to specify its functioning, as they determine the relationship between the input and output of the network [153]. The learning or adapting weights in a network are setting according to a training phase which requires data and is a major phase in developing an ANN. Various training procedures are required due to the structure of the networks [153].

Deep ANNs are composed of multiple processing layers in the network (Fig. 2.1b) so as to learn by representing data in different layers. The *layers* perform simple mathematical operations, such as convolution, pooling, and normalisation, which are implementable in a biological system [25, 126]. Feed-forward and recurrent neural networks are popular architectures of ANNs used in deep learning. A feed-forward neural network has a hierarchical network, that is ordered into layers (Fig.2.1b), where the first layer is the input layer, which that receives the data, the last layer is the output layer, which returns the results of network computations, and the layers in between are so-called hidden layers which mediate between the input and the output layers [147]. A recurrent neural network is a model that accounting time in the network by including feedback to the network at the local or global level. It allows the information to flow back from the output towards the input field [54, 63]. An idea behind the recurrent

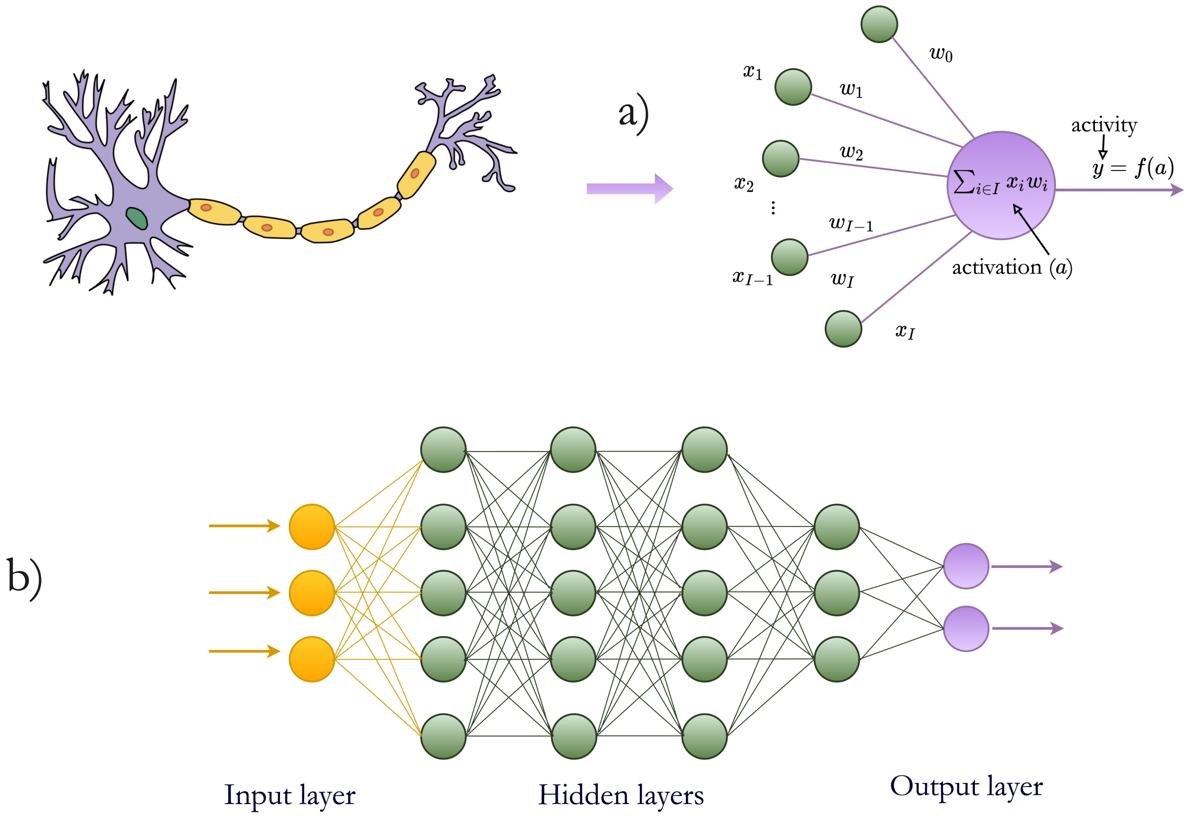


Fig. 2.1: Neurons are basic units of artificial neural networks, and learning is the propagation of signals through their connections, which interpolate as an automated weight change in the network. Deep learning, is an architecture of deep ANN which has several hidden layer for processing the input data before sending them to output layer for final processing and giving the output. The neuron illustration is taken from [17]

neural network is that to solve sequential or time-series problems, at some point the "current" input may not be sufficient and may requires access to previous information to solve them correctly [97]; therefore the current input is processed based on past as well as future inputs [54]. Hopfield network and in general associative memories are a class of recurrent neural networks (see Section 2.2.2).

In recent years, ANNs have rapidly progressed in different areas such as medicine, psychology, and neuroscience, in terms of meeting many challenges, such as image visualisation and segmentation, classification tasks, diagnosing disease, drug personalisation, and modelling cognitive behaviours [90, 98, 99, 126, 164].

2.2.2 *Neural associative memory*

Associative memory. Neural associative memory is a particular class of neural networks capable of memorising a set of patterns and with the ability to retrieve an original stored pattern from a noisy version or a partial clue [79, 99]. Associative memory is inspired by a similar concept in neuroscience, and as in artificial neural networks, it includes the representation of neurons. The term *associative* refers to the connection between two or more pieces of information when stored in the memory [79, 99]. For example, when you see a scene you may remember sequences from the movie, when you hear a literary phrase you may recall a poem with those words, or when you smell some food you may remember a journey.

Hopfield's network [65] was a milestone in the construction of the first auto-associative memory in 1982 (see Fig. 2.2a) [65]. Learning from the dynamical and circuit properties of neurons and their interactions, Hopfield was motivated to design a content-addressable memory to model this dynamic biological behaviour, and thus suggested a larger and more complex construction for a computer memory with a greater capacity [65]. Associative memories in general propose some beneficial properties such as robustness against noise, error correction, categorisation, storage capacity, and retrieval performance [65, 99].

Clique-based neural network. In brain activity, at any given time, a few neurons are firing simultaneously among an enormous population, which has motivated the design of more efficient associative memories [52, 53]. A clique-based network has a larger storage capacity and greater robustness in storage and retrieval compared with the Hopfield network [52]. Remember that a clique in a graph (network) is a complete sub-graph, for example, in Fig. 2.2 b) the red and green sub-graphs are two cliques of size four. In a clique-based neural network, neurons are split up into clusters. Each clique is formed by connecting neurons in the individual clusters [52, 53] (See Fig 2.2 b).

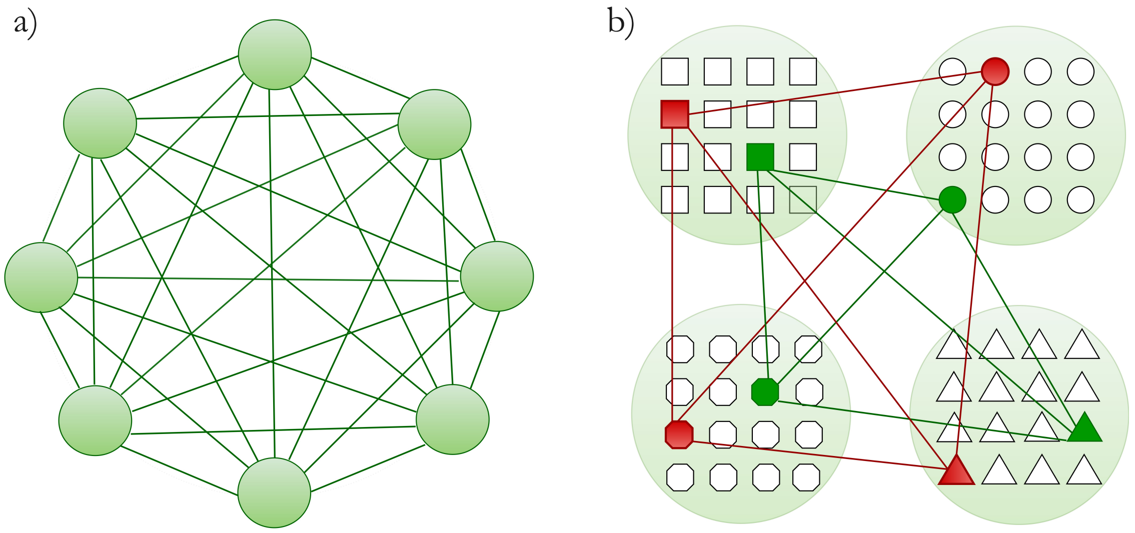


Fig. 2.2: A Hopfield neural network (a) is a complete graph, while a clique-based neural network (b) is a sparse network, with neurons split up into clusters, and the connections building cliques where each node is in one cluster.

Tournament-based neural network. Since anticipation and the forward direction of time are fundamental properties of human intelligence, considering the time and order of sequences during learning and storing in neural networks is an important factor for in imitating the brain network. Tournament-based neural network [72] are an extension of clique-based neural networks that have oriented clockwise connections (see Fig. 2.3), and therefore gains the ability to store sequential data as a chain of tournaments [91]. In graph theory, a tournament is a complete graph with directions assigned to all edges. This model is able to store sequences with a high degree of efficiency, where its mechanism of anticipation makes it more biologically plausible [72, 99]. In paper D [99] we proposed a more general tournament-based neural network that has some counterclockwise connection as well.

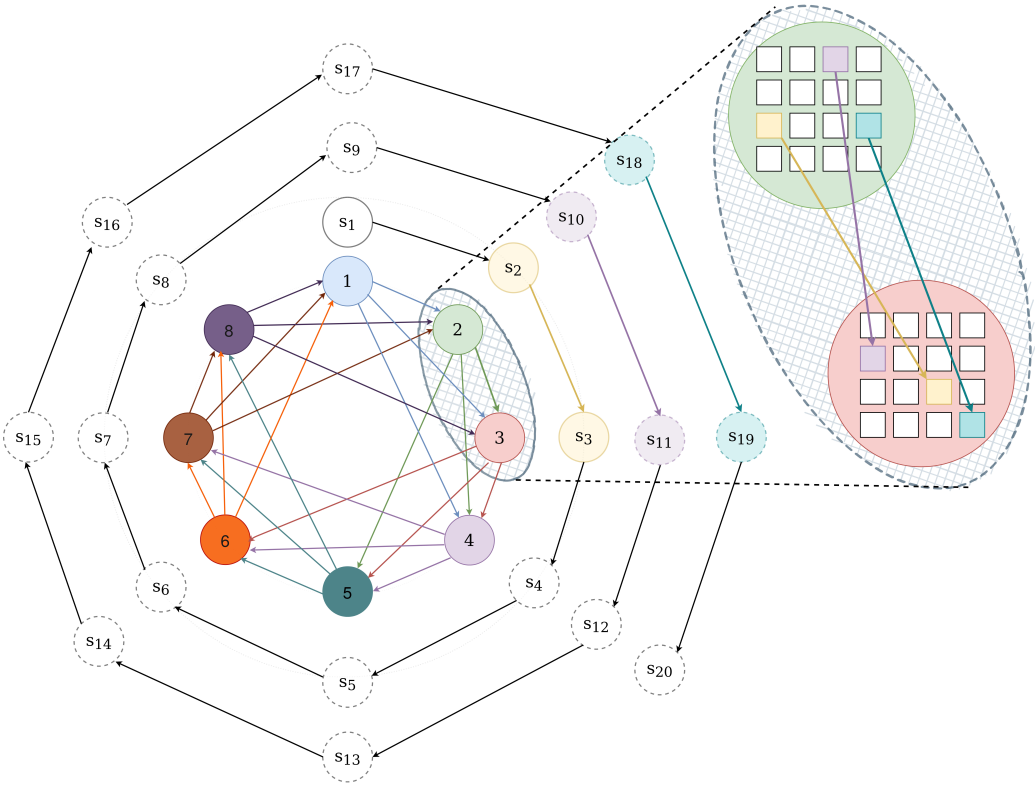


Fig. 2.3: Here is an illustration of a chain of tournaments for storing and retrieving a sequence with an arbitrary length [99]. The coloured circles are the clusters with the nodes inside them (see the enlarged clusters 1 and 2). The tournaments are size 4 (4 connected nodes), and an arrow represents a set of possible connections between nodes of two clusters (the arrow between clusters 1 and 2 represents three arrows illustrated in the zoomed in part). This network, for instance, allows storing a sequence of length 20, i.e. $s_1, \dots s_{20}$, that passes cyclically clusters and uses different nodes at each passage (the zoomed in part illustrate how this sequence passes three times through clusters 1 and 2). In this example, each node has arrows to three forward nodes, so giving the first three components (s_1, s_2 and s_3) to the retrieval algorithm retrieves the entire sequence. This figure is based on [72].

2.2.3 *Projective simulation*

Reinforcement learning Reinforcement learning refers to an area of machine learning concerned with how a learner (animals, human, or machine) learns what to do and how to map situations to actions by maximising the cumulative reward received from the environment (see Figure 2.4) [145]. The learner is in general not told which action to take, and must discover the correct action through trial and error. The feedback strategy in reinforcement learning is an evaluative process that deals with learning in sequential decision-making problems instead of correcting responses. Reinforcement learning is an interdisciplinary field usually studied in machine learning but also widely developed in other fields, especially in behaviour learning, psychology, neuroscience, and robotics [21, 98, 159].

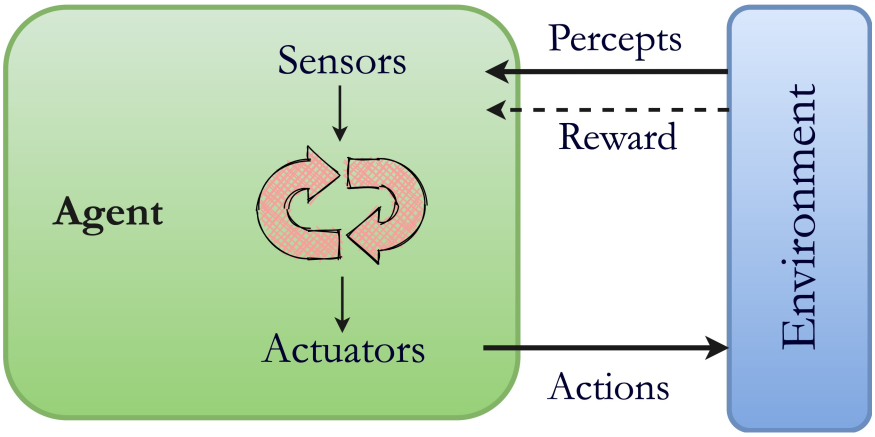


Fig. 2.4: Reinforcement learning is concerned with how a system learns by maximising the cumulative reward received from the environment. The learner discovers the correct action through a trial and error. This figure is based on Fig. 1 in [15].

Stimulus equivalence Stimulus equivalence attempts to explain how dissimilar events or ideas, particularly those that have never been related directly, are treated similarly. For instance, how might ideas A and B, which are both related to idea C, come to be related when they are not related directly and therefore all three ideas be used interchangeably [50]. Based on equivalence relation in mathematics, stimulus equivalence can be determined by properties of reflexivity (A = A), symmetry (if A = C then C = A), and transitivity (if A = C and B = C, then A = B) [137].

Stimulus equivalence was originally introduced by psychologists to understand the complexity of human behaviour and optimise teaching methodologies for children and adults with disabilities [2, 136] A number of groups of children with autism spectrum disorder and Down's syndrome [2], as well as adults with neurocognitive disorders such as mild cognitive impairment can benefit from these efficient learning techniques [3].

Projective simulation. Projective simulation is a new, physics-based approach to artificial intelligence that can connect the field of quantum physics to reinforcement learning. It can be seen as a type of reinforcement learning, with applications in advanced robotics as well as behavioural analysis [15, 94, 100]. Projective simulation is a system that is in continuous interaction with its environment so as to learn by trial and error through feedback; however, it has a more general framework than reinforcement learning, and can be applied to more problems, such as in quantum mechanics. A projective simulator resembles the internal representation in episodic memory and allows the agent (system) to project itself into a potential future, on the basis of previous experiences, and them make a new action [15] (see Fig.2.5). The episodic memory is a directed weighted network of episodes embedded in the projective simulator, that can be described as a stochastic network. The stochastic network instead of determining the next step for the agent by only the history of experiences, assigns the probabilities to the possible steps. The episodic memory is the most important difference between projective simulation and other standard reinforcement learning methods and allows the simulation of more complex features such as the future actions of agents. Projective simulation works on the basis of a random walk through the episodic memory network and updates it by creating new episodes and changing network weights [15] where an episode refers to a patch of stored previous experience. The structure of episodic memory in a projective simulation means this framework for reinforcement learning better resembles the real functionality of the brain [61, 101].

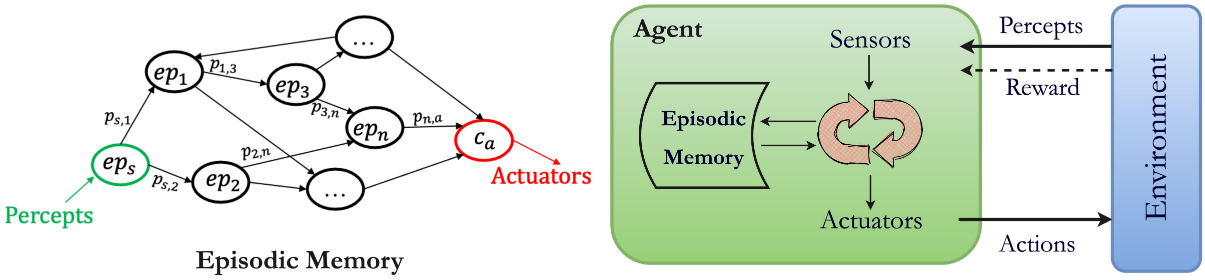


Fig. 2.5: Projective simulation [15] is a type of reinforcement learning with an embedded episodic memory. This model is based on a random walk through episodic memory to find the more probable actions. The episodic memory has a structural-dynamical property that enables the agent to detach the immediate connection with the environment and react upon its future exploration. This figure initially presented in Fig. 1 and 2 by Briegel et al [15].

Equivalence projective simulation. Stimulus equivalence and projective simulation are both used to study the complexity of behaviour. The former can be found in human subjects and the latter in artificial agents [100]. Equivalence projective simulation, as proposed by Mofrad et. al [100] is a novel machine learning model that can replicate human behaviour in terms of stimulus equivalence and which links the field of equivalence theory in behaviour analysis to an artificial agent in the machine learning area. Equivalence projective simulation updates the internal episodic memory of the original projective simulation to model several stimulus equivalence experiments. The learning process can be perceived in the re-configuration of the episodic memory by updating the weights, adding new episodes and new links.

This model derives new equivalent relations, symmetry, transitivity, and equivalence without receiving feedback from the environment. The model incorporates the ability, as in human memory, to forget while learning, and it is possible to model disabilities of memory by manipulating the parameters. Therefore, this model can simulate various behaviours such as the formation of equivalence relations in participants and the non-formation of equivalence relations in language-disabled children measured by real experiments [100].

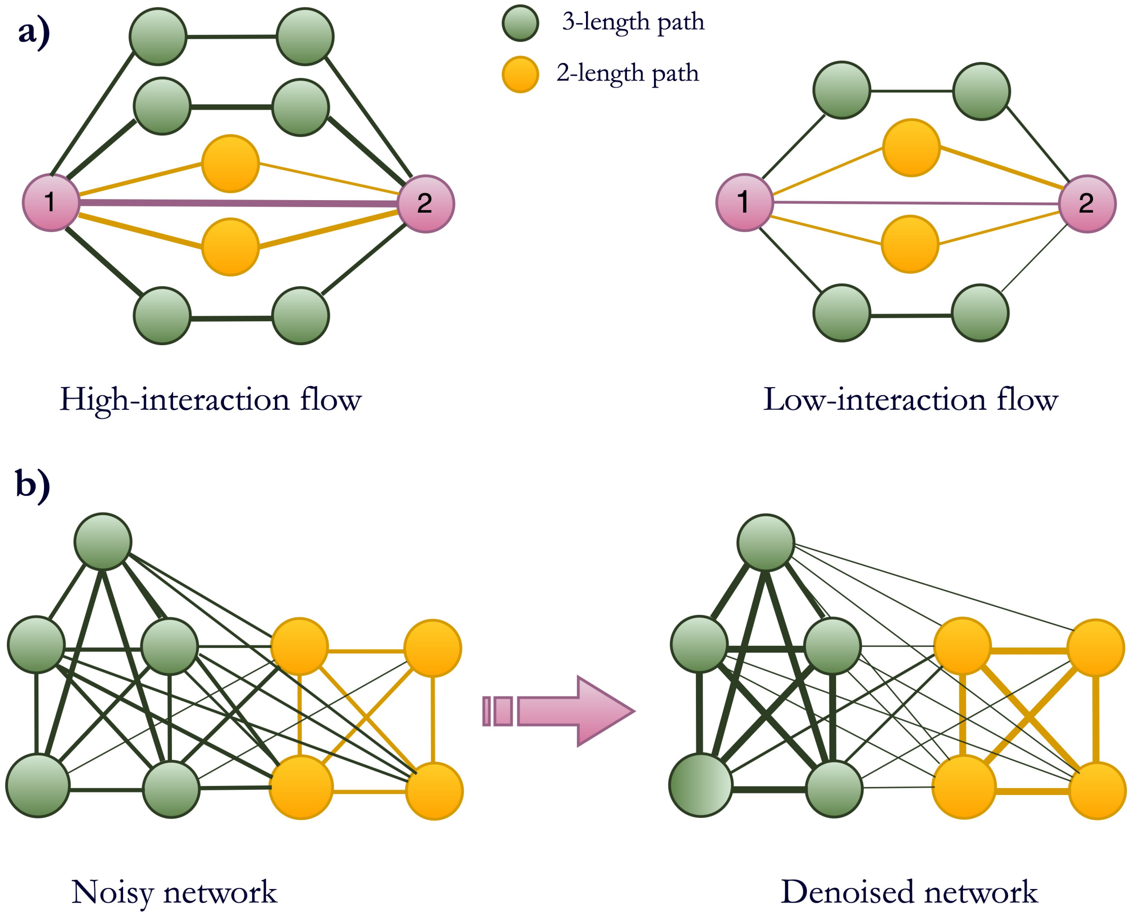


Fig. 2.6: The network enhancement updates a given weighted network with an iterative diffusion process. a) For updating the weight between any two nodes (e.g. node 1 and 2), all weighted paths between them with a length of three (green paths) or less (yellow and pink paths), are considered. The network on the left is an example of the strong connection between nodes 1 and 2, while the network on the right is an example of the weak connection between these two nodes. b) The network enhancement has an iterative process for updating the weights of the network. If there are strong paths between two nodes or if their connections are supported with many weak edges, the edge between them updates by strengthening the weight of the direct connection. It also weakens the edges' weight if they are not supported by many strong paths. This figure is based on Fig.1 in [155].

2.2.4 *Network enhancement*

Networks are prevalent in biological systems and encode the patterns of connectivity in their structures. Due to the complexity of biological organisation and the limitations of measurement technology, biological networks are noisy and unreliable. The noise in the network can affect the performance of the analysis by affecting the entire structure of the network, such as by changing the weight of edges that hide real biologically important connections. [66, 155].

Network enhancement is a diffusion-based computational method for denoising the biological network. It converts a noisy, undirected, weighted network into a new network with the same nodes but different weights for edges. The idea is that if nodes are connected through a path with high-weight edges, these nodes are more likely to be connected directly. This model uses random walks with a maximum length of three and generates a network by revising the weight of the edges. It removes the weak edges and enhances the edges with a high-weight path (see Fig. 2.6).

A network enhancement can be applied for denoising different networks by updating their connections. In paper E [101] a network enhancement model was applied to update the network of episodic memory, created by a projective simulation model [101].

"I never assume that we can argue from the truth of singular statements to the truth of theories. I never assume that by force of 'verified' conclusions, theories can be established as 'true', or even as merely 'probable'."

—Karl Popper [120]

CHAPTER 3

LONGITUDINAL DATA ANALYSIS AND MACHINE LEARNING

3.1 Longitudinal data

Data and data analysis improve our understanding of the world and help us to detect the consequences of events. Data can take various forms: numerical data, which has natural numerical values; categorical data, which counts the number of observations in each category, such as the number of females and males; cross-sectional data, which consist of observations measured at the same point in time; sequence of data, which is an enumerated collection of elements that occur or are arranged in a particular order; time-series, which contain observations of a sequence of points in time; and longitudinal data, which involves repeated observations of the same items at different points in time [140].

More broadly, longitudinal data can be understood as any information that tells us what has happened in a set of study cases over a sequence of time points; thus it is multi-dimension and multi-variate data involving measurements over time [36]. Sequences, time-series, and cross-sectional data can be seen as special cases of longitudinal data that are in one dimension only [33, 43].

A cohort study is a type of longitudinal study that engages and follows participants who share a common characteristic, such as a particular career, demographic similarity, or disease risk factor. By measuring outcomes during a follow-up period, it is possible to explore how and why the variables change. Cohort study is therefore an effective and robust method of establishing cause and effect [9]. It is an important method and among the most powerful approaches to research in the fields of epidemiology and bio-medicine, helping to explore and understand what factors affect the likelihood of developing a disease [9, 69]. Some diseases cause rapid fatality, making a study more difficult due to survivor bias. Cohort studies reduce the effect of this bias by involving factors in the study such as the calculation of incidence rates, relative risks, and confidence intervals [69, 86].

It is often possible to address the same questions in a longitudinal or cross-sectional study, but the major advantage of the former is its capacity to distinguish the effect of three time-related terms: cohort, period, and age [19, 33, 95]. Age effect show how individuals change as they age and progress through their lifespan, with variations often being linked to biological and social processes [10, 12]. Period effect show the changes that occur over a period of time, an external factors that equally affects all groups in the study, regardless of the age of the individuals. A cohort effect is a change characterising population as they move across time in a cohort (e.g. a unique experience) which is independent of the process of aging [12, 76]. The differences between age, duration, and cohort are demonstrated by Suzuki [146] in the following dialogue:

A: "I can't seem to shake off this tired feeling. Guess I'm just getting old. [Age effect]"

B: "Do you think it's stress? Business is down this year, and you've let your fatigue build up. [Period effect]"

A: "Maybe. What about you?"

B: "Actually, I'm exhausted too! My body feels really heavy."

A: "You're kidding. You're still young. I could work all day long when I was your age."

B: "Oh, really?"

A: "Yeah, young people these days are quick to whine. We were not like that. [Cohort effect]"

Fig. 3.1 illustrates a difference between cross-sectional and cohort study by an hypothetical example [33].

The benefits of longitudinal data are not without costs. In addition to some problems similar to those associated with cross-sectional studies, longitudinal designs have problems not included in cross-sectional designs, including problems with spurious measurement changes over time and missing data [51, 88, 95]. One issue with longitudinal measurements arises when distinguishing unreliability from true change, which can happens even if identical measurement instruments are used at different time-points. Differences in establishing study or lifespan changes may result in what is measured at one time-point being incomparable to what is measured at another [95].

In longitudinal data design, there are commonly missing observations. There is a risk of bias due to participants dropping-out of the study. For instance, in aging and health research, subjects may drop out of follow-ups due to out-of-scope residence, health, loss of interest in participating, or mortality [51, 87, 88]. Some studies describe a longitudinal study as "balanced" where data from repeated-measures has an equal number of observations for all variables in the study and as "unbalanced" where the number of observations is unequal. Missing data is possible in either type of design [87, 138].

Since, in a longitudinal study, the repeated measurements within one individual are correlated, analysing longitudinal data requires special statistical methods to draw valid scientific inferences. In addition, it is necessary to determine whether the data is balanced or unbalanced to indicate the appropriate statistical model [33, 87].

The most common techniques that researchers use to analyse longitudinal data are univariate methods, multivariate methods such as generalized estimating equations (GEE), and mixed-effects models [1, 8, 138, 150, 163].

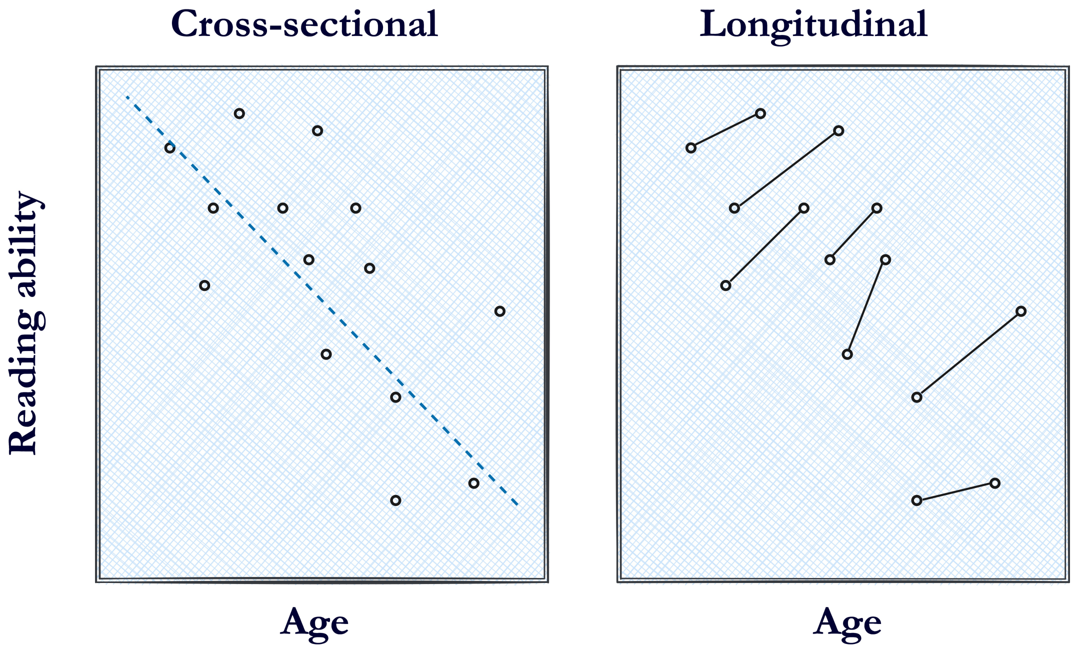


Fig. 3.1: This illustration compares the hypothetical cross-sectional and longitudinal studies originally presented in [33]. There can be entirely different results for the relationship between reading ability and age when taking cross-sectional or longitudinal approaches. In (a), the cross-sectional study shows a reduction in reading ability among older children. In (b), the same type of data was obtained from a longitudinal study with length two. This data shows that although the younger children began with a higher level of reading ability, everyone improved over time.

In univariate methods, the main aim is reducing the multiple outcomes into a single summary measure. In this approach, the observations for each individual are summarised as one value such as mean, median, maximum, last value, or slope of changes. These univariate measures are then compared across groups. By summarising measurements per person, the observation for the participants are independent, which means that various univariate techniques can be employed for analysing data. The approach is simple to understand and computation is easy. However, depending on the choice of summary function, there may be a substantial loss of information [1, 8].

Multivariate methods aim to estimate regression lines through data without explicitly modeling the correlation and covariance structures of the data. Traditional multivariate methods are used to test whether mean vectors in two or more samples have parallelism or differences in multivariate data. These methods can be used to analyse longitudinal data when observations are taken at the same time points for all subjects [1, 163]. Other important proposed methods for multivariate longitudinal data include linear models with correlated errors such as the generalised estimating equations (GEE) framework [163]. The GEE model is an extension of the generalised linear model from the independent data to the repeated longitudinal data setting [1]. GEE analysis estimates the regression coefficients in an iterative procedure to analyse the relationships between the variables in the model at different time points simultaneously. Indeed, GEE, by combining a within-subject relationship with a between-subjects relationship introduces a single regression coefficient for representing the data [150]. An extensive explanation of the details of GEE is beyond the scope of this work. Note that most of these multivariate methods do not handle large amounts of missing and irregularly spaced observations [1].

A mixed-effects model is one of the most natural solutions to model dependence among multivariate longitudinal data. It is a multivariate regression with explicit modelling of correlation and covariance structures, and can be thought of as a more general multivariate model. We explain mixed-effects models in some detail below (Section 3.2.1)

3.2 Predictive models

3.2.1 *Mixed-effects models*

Mixed-effects models are a flexible and powerful statistical tool in the analysis of group data such as longitudinal data in various areas. This method's flexibility in respect of modelling within-group correlations for both balanced and unbalanced longitudinal data in an identical framework has increased its popularity [119]. Mixed-effects models use a mixture of fixed effects – which refer to (1) regression coefficients that are associated with an entire population and are constant for all subjects and not allowed to be randomly varied at subject level and (2) random effects, which are associated with individuals to represent variability between-subjects and infer the conditional covariance structure [119, 138]. Mixed models thus perform estimation and inference on the regression coefficients in the data with multiple levels of grouping by considering within-subject correlation structures [8, 32, 138]. A random-effects approach can be used in linear, generalised linear, and non-linear mixed models [40] and such models are implemented in reliable and efficient statistical software [8, 119].

In this work we have used a linear mixed-effects model based on the model of West et al. [158] to estimate the values of each subject by considering the cohort effect in the data:

$$Y_{ij} = \underbrace{\beta_c X_{ij}^c}_{\text{fixed effect}} + \underbrace{u_{0i} + u_{1i} X_{ij} + \epsilon_{ij}}_{\text{random effect}}, \tag{3.1}$$

where Y_{ij} for subject $i : 1, \ldots, N$ (N is the number of subjects involved in the study) is the estimated value for the covariate that we modelled, at observation $j : 1, \ldots, n_i$ (n_i is the number of observations for subject i in unbalanced longitudinal data). X_{ij}^c for subject i is the predictor values obtained at time point j. $c : 0, 1, 2$ indicates the power of the predictor in the model. We have considered only one predictor in our models ($X = \text{age}$).β^c are fixed effect parameters whereas u_{0i}, and u_{1i} are random-effect parameters and ϵ_{ij} denotes random residual errors.

In this thesis (in Paper A and B) we have used linear mixed effects models to extract features from multivariate longitudinal data for later use for classification.

3.2.2 Classical machine learning

Machine learning consists of a set of tools and techniques based on computer science, statistics, and data sciences. These are algorithms that help computer systems improve automatically through experience [74]. Machine learning has been widely used in many fields of study, for clustering, classification, regression and more [122]. In this work (Papers A and B), we have used some well-known classification algorithms, which are matched with independent data, such as logistic regression, support vector machine, K-nearest neighbors, random forest, and a gradient boosting model, to propose predictive frameworks appropriate for longitudinal data [102, 103].

In the field of machine learning, a classification task has two meanings. The first is discovering existing classes within a set of observations; the other is when we know about certain classes and seek to set rules to classify a new observation as belonging to one of those classes. The former is referred to as unsupervised classification or clustering, and the latter as classification using supervised learning [96]. The term "classification" in our work refers to the one based on supervised learning.

3.2.3 Deep learning and convolutional neural networks

Artificial neural networks (discussed in Section 2.2.1), or deep neural networks, are a class of machine learning models that have recently outperformed other models in multiple tasks, particularly as regards image processing, cognitive tasks, and natural language processing [89, 122].

Convolutional neural networks, which have emerged from the study of the brain's visual cortex, are a specific architecture of artificial neural networks that are broadly used for analysing images, e.g. in computer vision [47]. The term "convolution" refers to the mathematical operations that process some of the filtering of the connections through the layers of the neural network to make its architecture more similar to that of real vision [47].

One of the challenges in training a neural network is over-fitting. In general, over-fitting occurs when the model performs well on the training set but poorly on the test set [47, 49]. To reduce the risk of over-fitting, different techniques are employed to regularise the tasks of neural networks. These techniques contain both implicit regularisation, such as data augmentation and transformation methods, and explicit regularisation, such as dropout, batch normalisation, and weight decay [35, 44, 49]. Our work in Paper C makes use of convolutional neural networks and various regularization techniques to perform an image classification task.

"Without big data, you are blind and deaf and in the middle of a freeway."

—Geoffrey Moore [106]

CHAPTER 4

DATA AND SOFTWARE

4.1 Data and ethics

The data used in this work was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI study was launched in 2003 as a public-private partnership, led by principal investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD [45]. Informed consent was obtained from all subjects prior to enrollment. All methods were carried out in accordance with relevant guidelines and regulation. The present study was approved by the ADNI Publication Committee (ADNI DPC).

We have also used data collected by the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) database (https://aibl.csiro.au). Launched in 2006, AIBL is the largest study in Australia to discover the biomarkers, cognitive characteristics, and health and lifestyle factors that determine the development of symptomatic AD. It comprises more than 1000 participants with a minimum age of 60 years and contains healthy volunteers and MCI and AD subjects. AIBL study methodology has been reported previously by Ellis et al. [37].

From these data sources, we selected longitudinal subjects who had at least two brain MRI scans, over a period of 15 years. Our data collection contains 1673 subjects (with a total of 8002 scans) from ADNI (7764 scans from 1603 subjects) and AIBL (238 scans from 70 subjects). In addition, for part of the work with these subjects, where available, we took their scores on three cognitive tests during the ADNI study: ADAS-cog-13, RAVLT-Immediate and RAVLT-Percent-Forgetting. In addition to the MRI scans and cognitive tests, we took gender, educational level and age into account in the analysis in this part. The subjects in this part had three types of cognitive diagnosis: cognitively normal (CN), mild cognitive impairment (MCI), and dementia or Alzheimer's Disease (AD).

Based on the ADNI labels, we defined four new subgroups, with a restriction to subjects with MRI scans at a minimum of two time-points. We labelled subjects as healthy controls (HCs) if they were labelled as CN at all visits, converted MCI (cMCI) if they were CN at their first visit but later converted to MCI during the study, stable MCI (sMCI) if they were MCI at all visits, and converted AD (cAD) if they were labelled as MCI at the first visit but later converted to AD (see Table 4.1). This differentiated the subjects according to stable and progressive groups, and led to the identification of characteristics associated with the rate of changes [104].

Labels in ADNI and our longitudinal labels		
Labels	Subgroup	Description
ADNI	CN	Cognitively normal at visit
	MCI	Mild cognitive impairment at visit
	Dementia	Alzheimer's disease at visit
This work	HC	CN at all visits
	cMCI	Initially CN but later converted to MCI
	sMCI	MCI at all visits
	cAD	Initially MCI but later converted to AD
	sAD	Dementia at all visits

Table 4.1: The original ADNI labels and the longitudinal labels used in the present study [103].

Collecting data for building a longitudinal study is an expensive and time-consuming process, with multiple pitfalls associated to the inherent challenges of longitudinal studies (see Section 3.1). Existing, large-scale and well-organized data collections such as ADNI and AIBL are therefore extremely valuable when developing and testing new methods. While this data has several advantages such as multimodality in measurement and a large number of participants, it has some limitations as well, which cannot easily be overcome as the data is already collected and annotated. One limitation is related to the disease diagnosis. The patients with MCI are highly diverse [27, 154], and clinical diagnosis of AD is uncertain [5]. Some studies have shown the AD is established in the brain years before impairments appear in the cognition [71]. Furthermore, in the datasets used in our work, the time between visits for subjects is on average half a year. It makes distinguishing between MCI and AD difficult as several patients are assumed as MCI, while AD has been established in their brain already. Furthermore, variability in MRI measurements, due to using scanners with different strengths and other environmental conditions during MRI examinations such as different head positions and head motion during scan, are some of the possible sources of instability in the data [31, 148].

4.2 Software libraries

FreeSurfer FreeSurfer is a powerful a widely used software package providing automated, robust, and accurate analyses of structural and functional human neuroimaging data [42]. The main focus of FreeSurfer is on analysing structural MRI scans of the brain, including segmentation of brain structures, segmentation of hippocampal sub-fields, segmentation of white matter by using diffusion MRI, and the reconstruction of surface models of the cerebral cortex. FreeSurfer is designed for the processing of both cross-sectional data and longitudinal data series [42, 124].

To obtain brain region volumes, we reprocessed all the T1-weighted MRI images used for our studies using the same version of FreeSurfer v.6.0 (the latest version at the time of the experiment) with the same systems: Ubuntu 18.04 GNU/Linux workstations. This was a very time-consuming process, but also necessary to reduce some of the variations we observed between the results obtained by different versions of FreeSurfer [103].

Python libraries. The code developed in this thesis was built on the basis of various libraries and functions in Python, especially `Pandas` [112], `Numpy` [59], `Matplotlib` [70], and `seaborn` [157]. In addition, in Papers A and B, we used the function `mixedlm` in the `statsmodels` library [134], version 0.11.0. to construct and fit the linear mixed-effects model. For the machine learning parts we used functions from `scikit-learn` [115] and `ELI5`. In Paper C, we employed `fastai` [67], a deep learning library based on `PyTorch` [114]. In Paper D, we used `NetworkX` [55] which is a library for analysing networks in Python. For drawing the figures we used `https://app.diagrams.net/`.

"The beauty of mathematics only shows itself to more patient followers."

—Maryam Mirzakhani [29]

CHAPTER 5

SUMMARY OF STUDIES

This chapter contains a summary of the five papers included in the thesis. Papers A, B, and C forms Part I of our work, and Paper D and E forms Part II.

5.1 Paper A: A predictive framework based on longitudinal trajectories: Application to detection of Alzheimer's disease

The main purpose of study A [103] is tackling balanced and imbalanced longitudinal data, containing noise and missing data at some time points, and to use them for prediction by applying standard machine learning methods that are not designed specifically for longitudinal data. We proposed a rather simple framework that uses mixed-effects models for extracting features from complex sets of longitudinal data together with well-known machine learning methods.

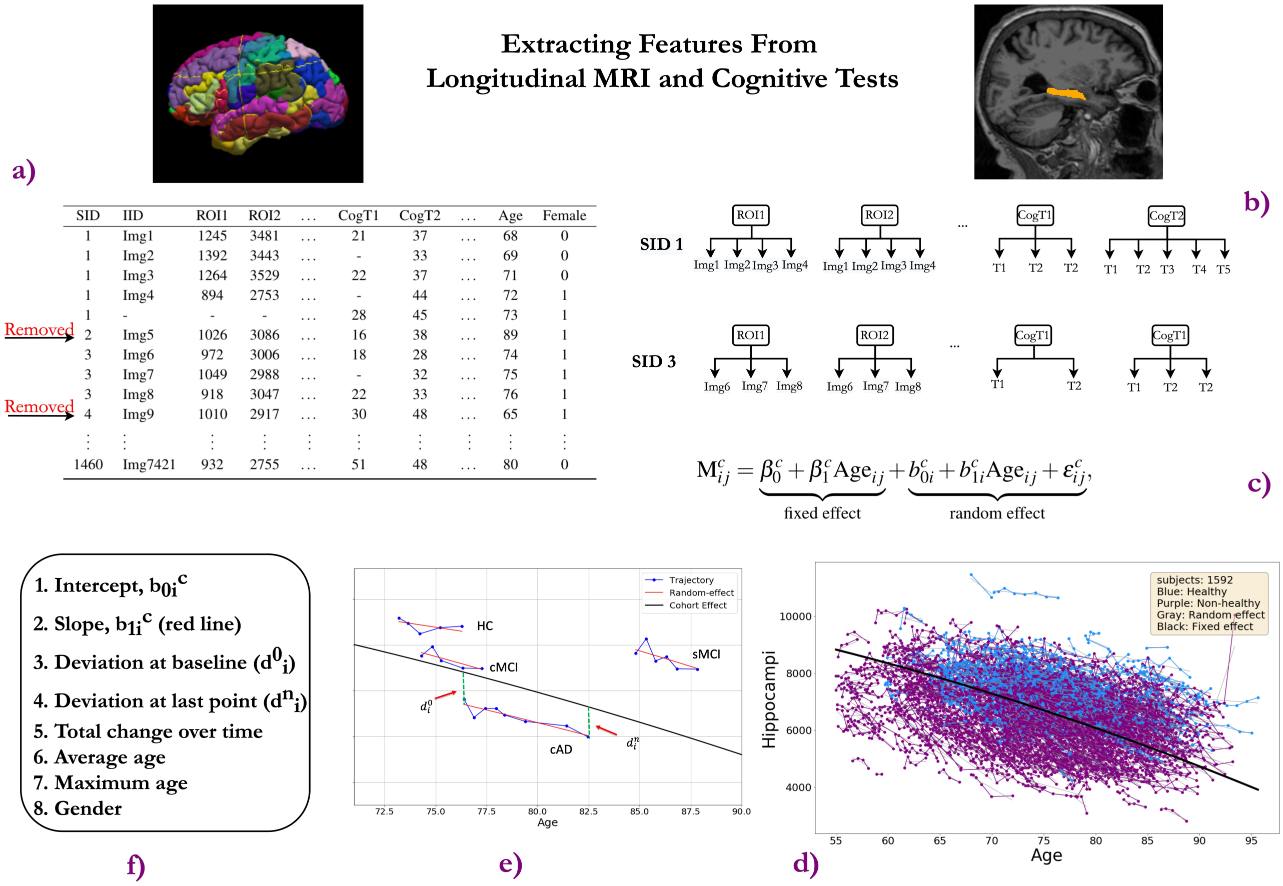
The main application of this study was the prediction of levels of dementia before clinical diagnosis, based on measuring volume atrophy in different regions of the brain. We applied our proposed framework to a collection of longitudinal brain MRI data from ADNI and AIBL for two predictive tasks (see data details in [103] and Section 4.1):

1. HC subjects vs. converted to MCI subjects

2. stable MCI subjects vs. converted to AD subjects

As the tasks were predicting future diagnostic status and investigating whether the model may indicate the risk of conversion, all information from the point of conversion onwards was removed. In other words, in task 1, with regard to cMCI subjects, we removed MRI scans that corresponded to clinical diagnoses of MCI, and in task 2, with regard to cAD subjects, we removed MRI scans that corresponded to AD. Fig. 5.1 illustrate the preparation steps for the features. Following the computation of brain region volumes using FreeSurfer on T1-weighted MRI recordings, a linear time-dependent mixed-effects model was used to derive features from brain volume trajectories. With the use of mixed-effects models, the instability and fluctuation observed in the volume trajectory become less influential (Fig. 5.1c, blue trajectory vs. red line), which leads to more robust features. For each variable, i.e. the hippocampus and ventricle volumes, we derived the characteristics of the slope of changes and the deviation from the cohort value for all subjects. In addition, we used the subjects' sex, average age at the time of scans, and the age at last scan (see Fig. 5.1d (1-4)).

Next, we used a soft voting strategy to train an ensemble model based on a logistic regression and a support vector machine. We constructed the model using data from the ADNI dataset, and evaluated it using data from non-overlapping subjects sourced from both ADNI and AIBL. On separate test sets, the model predicted the risk of MCI with an average accuracy of 69% and the risk of AD with an average accuracy of 75% ahead of the corresponding clinical diagnoses. The results indicate the ability of this framework to make early predictions of MCI and AD, before clinical diagnosis, based on volume atrophy in the brain.



Extracting Features From Longitudinal MRI and Cognitive Tests

SID	IID	ROI1	ROI2	…	CogT1	CogT2	…	Age	Female
1	Img1	1245	3481	…	21	37	…	68	0
1	Img2	1392	3443	…	-	33	…	69	0
1	Img3	1264	3529	…	22	37	…	71	0
1	Img4	894	2753	…	-	44	…	72	1
1	-	-	-	…	28	45	…	73	1
2	Img5	1026	3086	…	16	38	…	89	1
3	Img6	972	3006	…	18	28	…	74	1
3	Img7	1049	2988	…	-	32	…	75	1
3	Img8	918	3047	…	22	33	…	76	1
4	Img9	1010	2917	…	30	48	…	65	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1460	Img7421	932	2755	…	51	48	…	80	0

$$\mathrm{M}_{ij}^{c} = \underbrace{\beta_0^c + \beta_1^c \mathrm{Age}_{ij}}_{\text{fixed effect}} + \underbrace{b_{0i}^c + b_{1i}^c \mathrm{Age}_{ij} + \varepsilon_{ij}^c}_{\text{random effect}},$$

1. Intercept, $b_{0i}{}^{c}$
2. Slope, $b_{1i}{}^{c}$ (red line)
3. Deviation at baseline ($d^{0}{}_{i}$)
4. Deviation at last point ($d^{n}{}_{i}$)
5. Total change over time
6. Average age
7. Maximum age
8. Gender

Fig. 5.1: After segmentation of T1-weighted images with FreeSurfer, we had a table of volumes associated with brain regions. We combined this table with scores of some selected cognitive tests (in Paper B) for the same subjects. We kept participants with at least two MRI scans (a). For each subject, we had some selected regions of the brain (ROI) and cognitive test (CogT) at several time points except for missing data (b). We applied linear mixed effect model (c) to get the cohort regression and random effect for all subjects (d), separately for each variable (ROIs and CogTs). Next, we extracted the features from the results of the linear mixed models (e), and by adding some other information, we set up a vector of features for each individual (f), ready for applying machine learning methods.

5.2 Paper B: Cognitive and MRI trajectories for prediction of Alzheimer's disease

Paper B [104] is an extension of Paper A and has two main purposes. First, it assesses the application of the proposed framework in Paper A to different kinds of data. The experiment in Paper A included only longitudinal MRI data to predict MCI, whereas for Paper B, we used the same application as in Paper A but included cognitive measures and the following classification outcomes:

1. Classifying HC vs. cMCI
2. Classifying sMCI vs. cAD

Secondly, we investigated whether the prediction performance of the model improved when adding information from the MRI examinations.In this regard, we applied the pipeline from Paper A, a combination of mixed effects and machine learning models, to analyse a sample of multi-modal longitudinal data that include six sub-regions of the brain and performance on three cognitive tests. Morphometric brain measurements corresponding to the cognitive measures were selected. Brain measures associated with memory function (RAVL) included volumes of the entorhinal cortex and hippocampus. Furthermore, the total ADAS-Cog-13 score was used as a global measure corresponding to a global MRI volume measure of the lateral ventricles.

The features of the trajectories of change in cognitive and brain measures for the two pairs of subgroups (HC vs. cMCI), and (sMCI vs. cAD) were derived by applying statistical mixed-effects models (see Fig. 5.1 e-f). These features were then used to train an ensemble machine learning model to predict MCI and AD. The ensemble model was based on a soft voting strategy according to five models: logistic regression, support vector machine, K-nearest neighbours, random forest, and a gradient boosting classifier. To investigate which features were weighted highest in the classification tasks, we ran a permutation importance test to identify feature importance. The test is based on measuring the change in model accuracy when each feature is randomly shuffled multiple times. A feature is deemed to be more important than others if its permutation has a larger negative impact on the performance of the model.

We first applied the ensemble model to the features extracted from cognitive tests and then inspected whether the performance changed when adding the MRI features. Evaluation of the model in an independent test set indicated that the inclusion of MRI features substantially improved the classification of HC versus cMCI, while the result for sMCI versus cAD was only slightly improved. By integrating MRI features, the accuracy for (HC vs. cMCI) increased from 62% to 77%, whereas for (sMCI vs. cAD) changed from 77% to 78%. The results are in line with findings indicating that cognitive dysfunctions may become evident in a patient's performance on cognitive tests several years after Alzheimer's disease has established in the brain [71].

5.3 Paper C: From longitudinal measurements to image classification: Application to longitudinal MRI in Alzheimer's disease

In this study [102] we tackle the same problem in Paper A with a new approach that makes it possible to use complex sets of longitudinal data together with standard image classification methods. In this regard, we represented the longitudinal data from each subject as a two-dimensional grey-scale image and used them to train deep convolutional neural network image classifiers. Fig. 5.2 illustrates the setup.

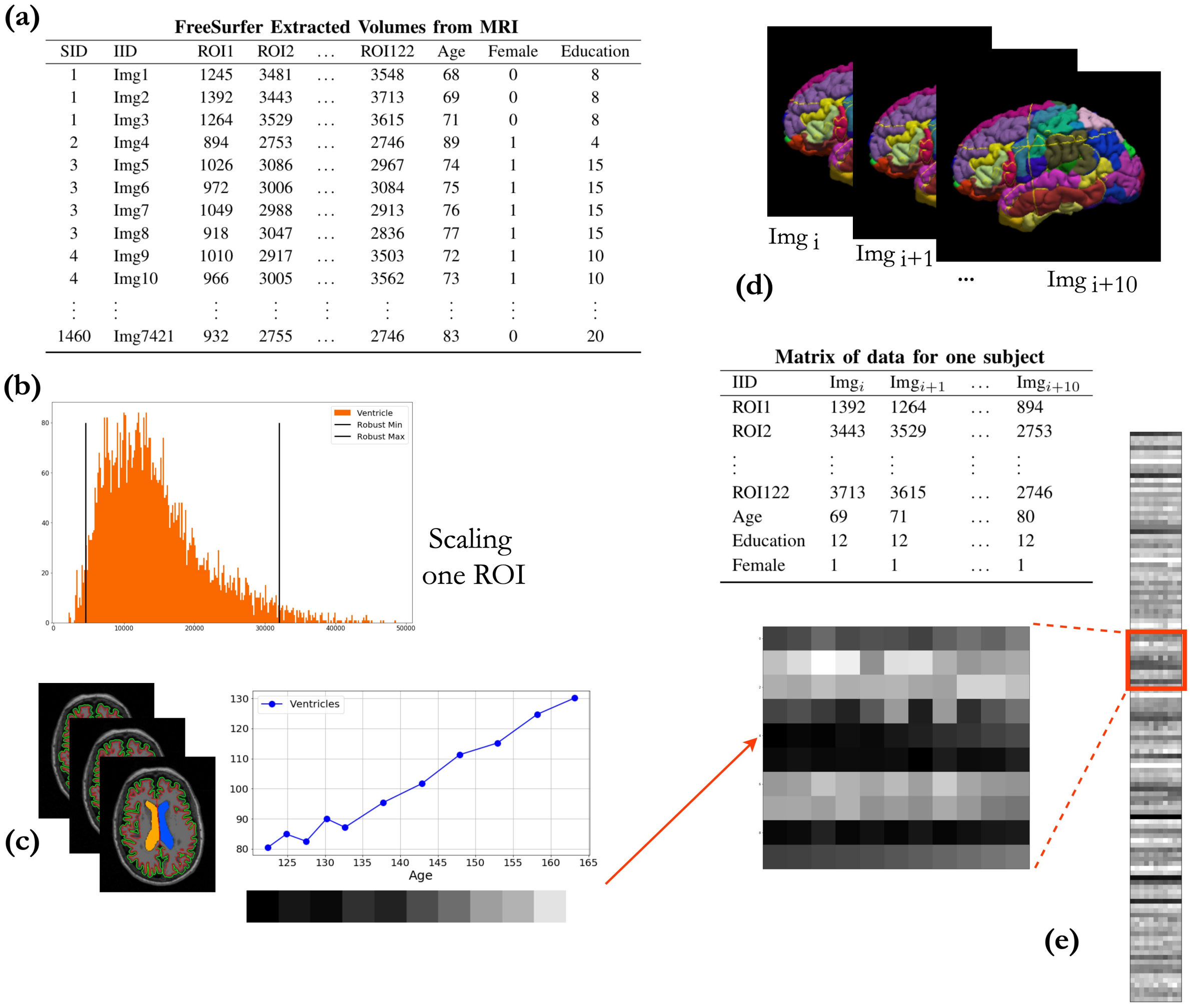
To evaluate our approach, we applied it to a set of longitudinal subjects (with at least three MRI scans) selected from the ADNI data source, containing ascending, descending, and categorical data. Our data set consisted of 736 subjects (female/male: 299/437) with a total of 3956 MRI scans. Note that the number of time points and the length between them varies significantly. The task was to classify two groups of subjects stable MCI versus converged AD (defined in Papers A and B)For each subject, we collected the measured volumes of all the regions in the brain using FreeSurfer 6.0 applied to the T1-weighted MR images, and we used the subjects' sex, level of education, and age at the time of the MRI examinations.

Before producing the two-dimensional images, the values associated with each feature, i.e. brain volume regions, age, gender, and education were scaled separately to obtain a standard range for each (Fig. 5.2: b). This was then used to scale the test set and other new previously unseen subjects. Then, for each subject, we gathered all the collected data in a matrix so that one axis was associated with time points and the other with the corresponding values of those time points (Fig. 5.2: d). Next, each scaled matrix was mapped onto a two-dimensional grey-scale image so that the pixel intensity represented the matrix values (Fig. 5.2: c and e). Note that before scaling we randomly selected the final test set. We then constructed a convolutional neural network for classifying sMCI versus cAD. During the training of our models, we used multiple regularisation techniques such as dropout, batch normalisation, and weight decay. Furthermore, to balance the class sizes and boost our models' generalisation ability, we augmented the data set by producing and adding Gaussian noise to the existent images. We conducted a grid search over the space of hyper-parameters to optimise their values and improve the model performance. Then we selected the top-performing models in terms of accuracy on validation data and calculated the final results by ensembling the selected models based on soft and hard voting strategies.

In an independent test set, we obtained average accuracies of 75.9% and 76.3% for hard and soft voting, respectively. This is a competitive result compared to other approaches for similar tasks, while the proposed technique is much simpler than other image classification methods using the original MRI recordings and can be applied to imbalanced longitudinal data.

5.4 Paper D: On neural associative memory structures: Storage and retrieval of sequences in a chain of tournaments

Learning and retrieval of temporal sequences in neural networks are fundamental properties of human intelligence. This study [99] proposed a structure for saving



FreeSurfer Extracted Volumes from MRI

SID	IID	ROI1	ROI2	...	ROI122	Age	Female	Education
1	Img1	1245	3481	...	3548	68	0	8
1	Img2	1392	3443	...	3713	69	0	8
1	Img3	1264	3529	...	3615	71	0	8
2	Img4	894	2753	...	2746	89	1	4
3	Img5	1026	3086	...	2967	74	1	15
3	Img6	972	3006	...	3084	75	1	15
3	Img7	1049	2988	...	2913	76	1	15
3	Img8	918	3047	...	2836	77	1	15
4	Img9	1010	2917	...	3503	72	1	10
4	Img10	966	3005	...	3562	73	1	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1460	Img7421	932	2755	...	2746	83	0	20

Matrix of data for one subject

IID	Img_i	Img_{i+1}	...	Img_{i+10}
ROI1	1392	1264	...	894
ROI2	3443	3529	...	2753
⋮	⋮	⋮	⋮	⋮
ROI122	3713	3615	...	2746
Age	69	71	...	80
Education	12	12	...	12
Female	1	1	...	1

Fig. 5.2: After preparing the table of brain volume extracted from T1-weighted images (a), we found a robust min and max for each ROI and replaced the upper and lower outliers with them (b). Then we scaled the volume in each column based on its robust min and max between 0 and 255. Therefore each longitudinal subject (with at least three scans) had scaled trajectories of volumes for all ROIs (c). For each subject, we presented a matrix that contains ROIs volume and age at all time points (here 11 points), in addition to gender and education level (d). This matrix was mapped to a two-dimensional gray-scale image so that pixel intensity represents the changes in the values (e). The images that can be input into image classification models.

a sequence, inspired by associative memory in the brain, and several methods for retrieving the whole sequence from corresponding incomplete versions of a previously stored sequence. The term *associative* here, as in brain studies, refers to the linkage of some pieces of information.

A tournament-based neural network (TNN) [72] is an associative memory that has directed clockwise connections, and therefore the ability to store sequential information (see Section 2.2.2 for the definition). In this paper, the TNN architecture was improved by proposing a more general structure for the learning procedure by adding some counterclockwise connections between elements of a sequence during learning. We refer to this new architecture feedback TNN.

This new structure supports retrieving a sequence, from any of its sufficiently large segments in two directions, regardless of where the segment is located. It makes the model more biologically plausible since it is known that the brain is able to follow the previously stored sequences, from any given point, both forwards and backwards (see e.g. [62]).

On the basis of human behaviour, two new retrieval techniques proposed in this study are have been called “Cache” and “Explore”. Cache retrieval reconsiders some previous randomly selected elements in a sequence in case an error is detected during the retrieval process. The idea behind the Cache technique can be simply depicted in a human decision-making procedure: imagine a person who quickly makes a decision, then if realises it is a mistake and tries to resolve it by revising past decisions. The results confirm that Cache techniques improved retrievals compare to the original retrieval algorithm. On the other hand, the Explore technique reduces the randomness in decisions by exploring the consequences of each decision while retrieving a sequence. Explore can be seen as a rather careful decision-maker who evaluates the consequences of all possible decisions at the time and then makes the best decision. Both Cache and Explore improved the results in terms of retrieving the correct sequence. Explore achieves the best results.

5.5 Paper E: Enhanced equivalence projective simulation: A framework for modeling formation of stimulus equivalence classes

This study was motivated by projective simulation [15] and equivalence projective simulation [100] models, which have an episodic memory that resembles the internal representation in the brain. Within these models, the agent can project itself into potential future situations before the real action is taken.

Study E [101] proposes an enhanced model for equivalence projective simulation [100]. This model can form and develop indirect connections during learning. In other words, the enhanced model is able to derive equivalence relations in a network without receiving feedback and information from the environment, and it links the field of equivalence in behaviour analysis to a machine learning context (i.e. a new type of reinforcement learning).

In the learning (training) phase, the derived relations are formed after completion of the training phase through an iterative diffusion process called network enhancement [155] (see Section 2.2.4). During this phase, the network structure (episodic memory) is updated and a noisy, indirect, weighted network is transformed into a new network with the same nodes but with updated connections and weights. This new memory is retrieved during the testing phase. The proposed model can be interpreted as resembling the brain’s learning procedure, which has formed connections and relations in its memory networks without directly being trained for them.

Although this model is of course far less complex than a brain neural network, it can simulate behaviour seen in some real experiments, including the formation of equivalence relations in typical participants and non-formation of equivalence relations in atypical groups such as among autistic children. Since the network enhancement is a denoising method, we can interpret the model as considering a typical memory as a less noisy memory, but a disabled memory as a noisy memory that cannot form equivalence relations.

In the model it is possible to adjust various factors such as learning rate, forgetting rate, formation of symmetry, and transitivity relations.The results of this approach are in line with recent findings in behavioural and neuroscience studies.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

The common core of this thesis is an exploration of memory and brain cognition. In Part I, the main goal is extracting the characteristic features from longitudinal data to be used together with standard machine learning techniques. In Part II, the main focus is on constructing computational models of memory and learning.

In Part I, we analyse neuroimaging data and psychometric tests to experimentally follow the changes in the brain and its functionality during aging and neurodegenerative disease. The predictive frameworks in this part allow us to derive biomarkers that detect alterations in brain volume and cognitive behaviour over time and to use standard machine- and deep learning techniques to predict the risk of dementia and highlight the model's most significant features. In the approach taken in Part II, brain networks and cognitive behaviour are modelled to improve knowledge about the connectivity and formation of relations in the brain, as well as to develop the technology of learning and retrieving in artificial memories, aiming to make them more efficient and realistic.

6.2 Future work

Some possible future directions for research based on this thesis include:

- Investigating the inclusion of further longitudinal data containing other types of neuroimaging sources such as functional MRI (fMRI), genetic profiles such as the APOE4 gene, and values from cerebrospinal fluid analyses, as well as other cognitive and biochemical measures in the studies of Part I.

- Applying the framework in Paper C to construct images from longitudinal fMRI time series that can then be used to predict interesting outcomes, e.g. dementia.

- Applying the frameworks in Part I to other kinds of data (in biomedical or non-medical fields) to check their efficacy for deriving features and making prediction based on longitudinal data.

- Investigating whether the diffusion network enhancement in Paper E, a method for denoising models for biological networks, can be applied to the adjacent matrices of functional MRI to reduce the noise, comparing their results with other denoising approaches.

- Developing the parameters of the model in Paper E for modeling memory disorder in patients with mild cognitive impairment and Alzheimer's disease by including longitudinal data from cognitive tests and fMRI.
- Taking the stimulus equivalence class presented in Paper E as a clique linking the stimuli. This will make it possible to connect the clique-based associative networks (Paper D) to the episodic memory in Paper E.

BIBLIOGRAPHY

[1] P. S. Albert. Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in medicine*, 18(13):1707–1732, 1999. 3.1

[2] E. Arntzen, L.-B. Halstadtro, E. Bjerke, K. J. Wittner, and A. Kristiansen. On the sequential and concurrent presentation of trials establishing prerequisites for emergent relations. *The Behavior Analyst Today*, 14(1-2):1, 2014. 2.2.3

[3] E. Arntzen and H. S. Steingrimsdottir. Electroencephalography (EEG) in the study of equivalence class formation. An explorative study. *Frontiers in human neuroscience*, 11:58, 2017. 2.2.3

[4] A. Association. 2019 Alzheimer's disease facts and figures. *Alzheimer's & dementia*, 15(3):321–387, 2019. 2.1.2

[5] A. P. Association. *Diagnostic and statistical manual of mental disorders (DSM-5)*. Pilgrim Press, Washington, 2013. 2.1.2, 4.1

[6] A. Baddeley. *Working memory, thought, and action*, volume 45. Oxford University Press, 2007. 2.1.1

[7] A. D. Baddeley. The influence of acoustic and semantic similarity on long-term memory for word sequences. *The Quarterly journal of experimental psychology*, 18(4):302–309, 1966. 2.1.1

[8] S. Bandyopadhyay, B. Ganguli, and A. Chatterjee. A review of multivariate longitudinal data analysis. *Statistical methods in medical research*, 20(4):299–330, 2011. 3.1, 3.2.1

[9] D. Barrett and H. Noble. What are cohort studies? 22(4):95–96. Publisher: Royal College of Nursing Section: Research made simple. 3.1

[10] A. Bell and K. Jones. Age, period and cohort processes in longitudinal and life course analysis: a multilevel perspective. *A life course perspective on health trajectories and transitions*, pages 197–213, 2015. 3.1

[11] C. Berrou, O. Dufor, V. Gripon, and X. Jiang. Information, noise, coding, modulation: What about the brain? In *2014 8th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*, pages 167–172. IEEE, 2014. 2.2

[12] R. D. Blanchard, J. B. Bunker, and M. Wachs. Distinguishing aging, period and cohort effects in longitudinal studies of elderly populations. *Socio-Economic Planning Sciences*, 11(3):137–146, 1977. 3.1

[13] P. Boyle, R. Wilson, N. Aggarwal, Y. Tang, and D. Bennett. Mild cognitive impairment: risk of Alzheimer disease and rate of cognitive decline. *Neurology*, 67(3):441–445, 2006. 2.1.2

[14] M. Brand and H. J. Markowitsch. The principle of bottleneck structures. In *Principles of learning and memory*, pages 171–184. Springer, 2003. 2.1.1

[15] H. J. Briegel and G. De las Cuevas. Projective simulation for artificial intelligence. *Scientific reports*, 2(1):1–16, 2012. 1.2, 2.2, 2.4, 2.2.3, 2.5, 5.5

[16] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi. Forecasting the global burden of Alzheimer's disease. *Alzheimer's & dementia*, 3(3):186–191, 2007. 2.1.2

[17] C. by Dhp1080. Derived Neuron schema with no labels.svg. `https://commons.wikimedia.org/wiki/File:Derived_Neuron_schema_with_no_labels.svg`. Accessed: 2021-06-01. 2.1

[18] J. S. W. Campbell and G. Bruce Pike. Diffusion Magnetic Resonance Imaging. In R. Narayan, editor, *Encyclopedia of Biomedical Engineering*, pages 505–518. Elsevier, Oxford, 2019. 2.1.3

[19] E. J. Caruana, M. Roman, J. Hernández-Sánchez, and P. Solli. Longitudinal studies. *Journal of thoracic disease*, 7(11):E537, 2015. 3.1

[20] A. Chandra, G. Dervenoulas, M. Politis, A. D. N. Initiative, et al. Magnetic resonance imaging in Alzheimer's disease and mild cognitive impairment. *Journal of Neurology*, 266(6):1293–1302, 2019. 2.1.3

[21] J. Choi, K. Park, M. Kim, and S. Seok. Deep reinforcement learning of navigation in a complex and crowded environment with a limited field of view. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5993–6000. IEEE, 2019. 2.2.3

[22] N. Chomsky. Language and thought. 1993. I

[23] E. T. Chou and J. A. Carrino. chapter 10 - Magnetic Resonance Imaging. In S. D. Waldman and J. I. Bloch, editors, *Pain Management*, pages 106–117. W.B. Saunders, Philadelphia, 2007. 2.1.3

[24] R. M. Cichy and D. Kaiser. Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317, 2019. 1.1

[25] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):1–13, 2016. 2.2.1

[26] J. Cleary and A. Guimarães. Magnetic Resonance Imaging. In L. M. McManus and R. N. Mitchell, editors, *Pathobiology of Human Disease*, pages 3987–4004. Academic Press, San Diego, 2014. 2.1.3

[27] J. H. Cole and K. Franke. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in neurosciences*, 40(12):681–690, 2017. 4.1

[28] J. R. Crawford, A. Venneri, and R. E. O'Carroll. 7.07 - Neuropsychological Assessment of the Elderly. In A. S. Bellack and M. Hersen, editors, *Comprehensive Clinical Psychology*, pages 133–169. Pergamon, Oxford, 1998. 2.1.4

[29] N. D. Daniel Mathews. The beauty of mathematics shows itself to patient followers. `https://www.danielmathews.info/2018/05/21/the-beauty-of-mathematics-shows-itself-to-patient-followers-the-work-of-ma` Accessed: 2021-06-01. 4.2

[30] N. A. Dennis, I. C. Turney, C. E. Webb, and A. A. Overman. The effects of item familiarity on the neural correlates of successful associative memory encoding. *Cognitive, Affective, & Behavioral Neuroscience*, 15(4):889–900, 2015. 2.1.1

[31] X. Di, M. Wolfer, S. Kühn, Z. Zhang, and B. B. Biswal. Estimations of the weather effects on brain functions using functional MRI–a cautionary tale. *bioRxiv*, page 646695, 2019. 4.1

[32] R. Diez. A glossary for multilevel analysis. *Journal of epidemiology and community health*, 56(8):588, 2002. 3.2.1

[33] P. Diggle, P. J. Diggle, P. Heagerty, K.-Y. Liang, P. J. Heagerty, S. Zeger, et al. *Analysis of longitudinal data*. Oxford University Press, 2002. 3.1, 3.1

[34] R. Dodel, A. Rominger, P. Bartenstein, F. Barkhof, K. Blennow, S. Förster, Y. Winter, J.-P. Bach, J. Popp, J. Alferink, et al. Intravenous immunoglobulin for treatment of mild-to-moderate Alzheimer's disease: a phase 2, randomised, double-blind, placebo-controlled, dose-finding trial. *The Lancet Neurology*, 12(3):233–243, 2013. 2.1.2

[35] N. Dvornik, J. Mairal, and C. Schmid. On the importance of visual context for data augmentation in scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3.2.3

[36] J. Elliott, J. Holland, and R. Thomson. Longitudinal and panel studies. *The SAGE handbook of social research methods*, pages 228–248, 2008. 3.1

[37] K. A. Ellis, A. I. Bush, D. Darby, D. De Fazio, J. Foster, P. Hudson, N. T. Lautenschlager, N. Lenzo, R. N. Martins, P. Maruff, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, 21(4):672–687, 2009. 4.1

[38] M. S. Fanselow and A. M. Poulos. The neuroscience of mammalian associative learning. *Annu. Rev. Psychol.*, 56:207–234, 2005. 2.1.1

[39] E. K. Fard, J. L. Keelor, A. A. Bagheban, and R. W. Keith. Comparison of the Rey Auditory Verbal Learning Test (RAVLT) and digit test among typically achieving and gifted students. *Iranian journal of child neurology*, 10(2):26, 2016. 2.1.4

[40] S. Fieuws and G. Verbeke. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431, 2006. 3.2.1

[41] K. R. T. Fink and J. R. Fink. 4 - Principles of Modern Neuroimaging. In R. G. Ellenbogen, L. N. Sekhar, N. D. Kitchen, and H. B. da Silva, editors, *Principles of Neurological Surgery (Fourth Edition)*, pages 62–86.e2. Elsevier, Philadelphia, fourth edition edition, 2018. 2.1.3

[42] B. Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. 4.2

[43] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied longitudinal analysis*, volume 998. John Wiley & Sons, 2012. 3.1

[44] C. Garbin, X. Zhu, and O. Marques. Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, pages 1–39, 2020. 3.2.3

[45] G. Gavidia-Bovadilla, S. Kanaan-Izquierdo, M. Mataró-Serrat, A. Perera-Lluna, A. D. N. Initiative, et al. Early prediction of Alzheimer's disease using null longitudinal model-based classifiers. *PloS one*, 12(1):e0168011, 2017. 4.1

[46] Y. E. Geda. Mild cognitive impairment in older adults. *Current psychiatry reports*, 14(4):320–327, 2012. 2.1.2

[47] A. Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019. 2.2.1, 3.2.3

[48] G. H. Glover. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics*, 22(2):133–139, 2011. 2.1.3

[49] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep Learning*. MIT Press, 2016. 3.2.3

[50] G. Green and R. R. Saunders. Stimulus equivalence. In *Handbook of research methods in human operant behavior*, pages 229–262. Springer, 1998. 2.2.3

[51] D. A. Grimes and K. F. Schulz. Cohort studies: marching towards outcomes. *The Lancet*, 359(9303):341–345, 2002. 3.1

[52] V. Gripon and C. Berrou. Sparse neural networks with large learning diversity. *IEEE transactions on neural networks*, 22(7):1087–1096, 2011. 2.2.2

[53] V. Gripon, J. Heusel, M. Löwe, and F. Vermet. A comparative study of sparse associative memories. *Journal of Statistical Physics*, 164(1):105–129, 2016. 2.2.2

[54] P. Gupta and N. K. Sinha. CHAPTER 14 - Neural Networks for Identification of Nonlinear Systems: An Overview. In N. K. Sinha and M. M. Gupta, editors, *Soft Computing and Intelligent Systems*, Academic Press Series in Engineering, pages 337–356. Academic Press, San Diego, 2000. 2.2.1

[55] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. In G. Varoquaux, T. Vaught, and J. Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008. 4.2

[56] J. Haile. Toward technical understanding: Part 2. Elementary levels. *Chemical Engineering Education*, 31(4):214–219, 1997. 2.1.1

[57] J. Haile. Toward technical understanding: Part 3. Advanced levels. *Chemical Engineering Education*, 32(1):30–39, 1998. 2.1.1

[58] S. Hammond. Using psychometric tests. *Research methods in psychology*, 3:182–209, 2006. 2.1.4

[59] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. 4.2

[60] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017. 1.1, 2.2.1

[61] M. E. Hasselmo. *How we remember: brain mechanisms of episodic memory*. MIT press, 2011. 2.2.3

[62] J. Hawkins and S. Blakeslee. *On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines*. Macmillan, 2007. 5.4

[63] S. Haykin. CHAPTER 4 - Neural Networks: A Guided Tour. In N. K. Sinha and M. M. Gupta, editors, *Soft Computing and Intelligent Systems*, Academic Press Series in Engineering, pages 71–80. Academic Press, San Diego, 2000. 2.2.1

[64] H. Hippius and G. Neundörfer. The discovery of Alzheimer's disease. *Dialogues in clinical neuroscience*, 5(1):101, 2003. 2.1.2

[65] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982. 2.2.2

[66] G. Hornung and N. Barkai. Noise propagation and signaling sensitivity in biological networks: a role for positive feedback. *PLoS Comput Biol*, 4(1):e8, 2008. 2.2.4

[67] J. Howard and S. Gugger. Fastai: A layered API for deep learning. *Information*, 11(2):108, 2020. 4.2

[68] S. A. Huettel, A. W. Song, G. McCarthy, et al. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA, 2004. 2.1.3

[69] S. B. Hulley. *Designing clinical research*. Lippincott Williams & Wilkins, 2007. 3.1

[70] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. 4.2

[71] C. R. Jack Jr and D. M. Holtzman. Biomarker modeling of Alzheimer's disease. *Neuron*, 80(6):1347–1358, 2013. 4.1, 5.2

[72] X. Jiang, V. Gripon, C. Berrou, and M. Rabbat. Storing sequences in binary tournament-based neural networks. *IEEE transactions on neural networks and learning systems*, 27(5):913–925, 2016. 1.3, 2.2.2, 2.3, 5.4

[73] D. K. Jones. *Diffusion MRI*. Oxford University Press, 2010. 2.1.3

[74] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015. 3.2.2

[75] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018. 1.1

[76] K. M. Keyes and G. Li. Age–period–cohort modeling. In *Injury research*, pages 409–426. Springer, 2012. 3.1

[77] S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014. 1.1

[78] J. H. King, J. D. Gfeller, and H. P. Davis. Detecting simulated memory impairment with the Rey Auditory Verbal Learning Test: Implications of base rates and study generalizability. *Journal of clinical and experimental neuropsychology*, 20(5):603–612, 1998. 2.1.4

[79] T. Kohonen. *Associative memory: A system-theoretical approach*, volume 17. Springer Science & Business Media, 2012. 2.2.2

[80] I. Kononenko and M. Kukar. Chapter 11 - Artificial Neural Networks. In I. Kononenko and M. Kukar, editors, *Machine Learning and Data Mining*, pages 275–320. Woodhead Publishing, 2007. 2.2.1

[81] I. Kononenko and M. Kukar. *Machine learning and data mining*. Horwood Publishing, 2007. 2.2.1

[82] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1.1

[83] J. K. Kueper, M. Speechley, and M. Montero-Odasso. The Alzheimer's disease assessment scale–cognitive subscale (ADAS-Cog): modifications and responsiveness in pre-dementia populations. a narrative review. *Journal of Alzheimer's Disease*, 63(2):423–444, 2018. 2.1.4

[84] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1.1

[85] R. L. Leong, J. C. Lo, S. K. Sim, H. Zheng, J. Tandi, J. Zhou, and M. W. Chee. Longitudinal brain structure and cognitive changes over 8 years in an East Asian cohort. *NeuroImage*, 147:852–860, 2017. 2.1.3

[86] D. E. Lilienfeld, D. E. Lilienfeld, P. D. Stolley, A. M. Lilienfeld, et al. *Foundations of epidemiology*. Oxford University Press, USA, 1994. 3.1

[87] X. Liu. Chapter 1 - Introduction. In X. Liu, editor, *Methods and Applications of Longitudinal Data Analysis*, pages 1–18. Academic Press. 3.1

[88] X. Liu. *Methods and applications of longitudinal data analysis*. Elsevier, 2015. 3.1

[89] A. J. Lundervold, A. Vik, and A. Lundervold. Lateral ventricle volume trajectories predict response inhibition in older age—A longitudinal brain imaging and machine learning approach. *Plos one*, 14(4):e0207967, 2019. 3.2.3

[90] A. S. Lundervold and A. Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019. 2.2.1

[91] M. R. S. Marques, G. B. Hacene, C. E. R. K. Lassance, and P.-H. Horrein. Large-Scale Memory of Sequences Using Binary Sparse Neural Networks on GPU. In *2017 International Conference on High Performance Computing & Simulation (HPCS)*, pages 553–559. IEEE, 2017. 2.2.2

[92] P. M. Matthews and P. Jezzard. Functional magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(1):6–12, 2004. 2.1.3

[93] L. E. Matzen, M. C. Trumbo, R. C. Leach, and E. D. Leshikar. Effects of non-invasive brain stimulation on associative memory. *Brain research*, 1624:286–296, 2015. 2.1.1, 2.1.3

[94] A. A. Melnikov, A. Makmal, V. Dunjko, and H. J. Briegel. Projective simulation with generalization. *Scientific reports*, 7(1):1–14, 2017. 2.2.3

[95] S. Menard. Longitudinal Studies, Panel. In K. Kempf-Leonard, editor, *Encyclopedia of Social Measurement*, pages 601–607. Elsevier. 3.1, 3.1

[96] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. Machine learning, neural and statistical classification. 1994. 2.2.1, 3.2.2

[97] B. R. Mitchell. Chapter 3 - Overview of advanced neural network architectures. In S. Cohen, editor, *Artificial Intelligence and Deep Learning in Pathology*, pages 41–56. Elsevier, 2021. 2.2.1

[98] A. A. Mofrad. When Behavior Analysis Meets Machine Learning; Formation of Stimulus Equivalence Classes and Adaptive Learning in Artificial Agents. 2021. 2.2.1, 2.2.3

[99] A. A. Mofrad, S. A. Mofrad, A. Yazidi, and M. G. Parker. On Neural Associative Memory Structures: Storage and Retrieval of Sequences in a Chain of Tournaments, 2021. Accepted. 1.2, 2.1.1, 2.2.1, 2.2.2, 2.2.2, 2.3, 5.4

[100] A. A. Mofrad, A. Yazidi, H. L. Hammer, and E. Arntzen. Equivalence projective simulation as a framework for modeling formation of stimulus equivalence classes. *Neural computation*, 32(5):912–968, 2020. 2.2.3, 2.2.3, 5.5

[101] A. A. Mofrad, A. Yazidi, S. A. Mofrad, H. L. Hammer, and E. Arntzen. Enhanced Equivalence Projective Simulation: A Framework for Modeling Formation of Stimulus Equivalence Classes. *Neural computation*, 33(2):483–527, 2021. 2.2, 2.2.3, 2.2.4, 5.5

[102] S. A. Mofrad, H. Bartsch, A. S. Lundervold, and A. D. N. Initiative. From longitudinal measurements to image classification: application to longitudinal MRI in Alzheimer's disease. *Under review*, 2021. 3.2.2, 5.3

[103] S. A. Mofrad, A. Lundervold, A. S. Lundervold, A. D. N. Initiative, et al. A predictive framework based on brain volume trajectories enabling early detection of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, page 101910, 2021. 3.2.2, 4.1, 4.2, 5.1

[104] S. A. Mofrad, A. J. Lundervold, A. Vik, and A. S. Lundervold. Cognitive and MRI trajectories for prediction of Alzheimer's disease. *Scientific Reports*, 11(1):1–10, 2021. 2.1.4, 4.1, 5.2

[105] S. A. Montgomery, L. Thal, and R. Amrein. Meta-analysis of double blind randomized controlled clinical trials of acetyl-L-carnitine versus placebo in the treatment of mild cognitive impairment and mild Alzheimer's disease. *International Clinical Psychopharmacology*, 18(2):61–71, 2003. 2.1.2

[106] G. Moore. Big data. https://en.wikiquote.org/wiki/Geoffrey_Moore. 3.2.3

[107] E. Moradi, I. Hallikainen, T. Hanninen, J. Tohka, A. D. N. Initiative, et al. Rey's Auditory Verbal Learning Test scores can be predicted from whole brain MRI in Alzheimer's disease. *NeuroImage: Clinical*, 13:415–427, 2017. 2.1, 2.1.4

[108] P. Müller and D. R. Insua. Issues in Bayesian analysis of neural network models. *Neural Computation*, 10(3):749–770, 1998. 2.2.1

[109] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 2.2.1

[110] R. Narayan. *Encyclopedia of biomedical engineering*. Elsevier, 2018. 2.1.3

[111] E. Nichols, C. E. Szoeke, S. E. Vollset, N. Abbasi, F. Abd-Allah, J. Abdela, M. T. E. Aichour, R. O. Akinyemi, F. Alahdab, S. W. Asgedom, et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, 18(1):88–106, 2019. 2.1.2

[112] T. pandas development team. pandas-dev/pandas: Pandas, Feb. 2020. 4.2

[113] D. C. Park and P. Reuter-Lorenz. The adaptive brain: aging and neurocognitive scaffolding. *Annual review of psychology*, 60:173–196, 2009. 2.1.2

[114] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 4.2

[115] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 4.2

[116] R. C. Petersen. Mild cognitive impairment or questionable dementia. *Archives of neurology*, 57(5):643–644, 2000. 2.1.2

[117] R. C. Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of internal medicine*, 256(3):183–194, 2004. 2.1.2

[118] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen. Mild cognitive impairment: clinical characterization and outcome. *Archives of neurology*, 56(3):303–308, 1999. 2.1.2

[119] J. Pinheiro and D. Bates. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2006. 3.2.1

[120] K. Popper. *The logic of scientific discovery*. Routledge, 2005. 2.2.4

[121] M. J. Prince. *World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease International, 2015. 2.1.2

[122] E. R. Ranschaert, S. Morozov, and P. R. Algra. *Artificial intelligence in medical imaging: opportunities, applications and risks*. Springer, 2019. 3.2.2, 3.2.3

[123] N. Raz. Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In F. Craik and T. Salthouse, editors, *The handbook of aging and cognition*. Lawrence Erlbaum Associates Publishers, 2000. 2.1.3

[124] M. Reuter and B. Fischl. Avoiding asymmetry-induced bias in longitudinal image processing. *NeuroImage*, 57(1):19–21, 2011. 4.2

[125] P. A. Reuter-Lorenz and C. Lustig. Brain aging: reorganizing discoveries about the aging mind. *Current opinion in neurobiology*, 15(2):245–251, 2005. 2.1.2

[126] B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019. 1.1, 2.2.1, 2.2.1

[127] K. M. Rodrigue and N. Raz. Shrinkage of the entorhinal cortex over five years predicts memory performance in healthy adults. *Journal of Neuroscience*, 24(4):956–963, 2004. 2.1.3

[128] E. T. Rolls and A. Treves. Neural networks in the brain involved in memory and recall. *Progress in brain research*, 102:335–341, 1994. 2.1.1, 2.1.3

[129] W. G. Rosen, R. C. Mohs, and K. L. Davis. A new rating scale for Alzheimer's disease. *The American journal of psychiatry*, 1984. 2.1.4

[130] M. Roser, E. Ortiz-Ospina, and H. Ritchie. Life expectancy. *Our World in Data*, 2013. 2.1.2

[131] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010. 2.1

[132] J. Sacramento, R. P. Costa, Y. Bengio, and W. Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. *arXiv preprint arXiv:1810.11393*, 2018. 1.1

[133] M. Schmidt et al. *Rey auditory verbal learning test: A handbook*. Western Psychological Services Los Angeles, CA, 1996. 2.1.4

[134] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*, 2010. 4.2

[135] L. Sherwood. *Human physiology: from cells to systems*. Cengage learning, 2015. 2.1.1

[136] M. Sidman, O. Cresson Jr, and M. Willson-Morris. Acquisition of matching to sample via mediated transfer. *Journal of the Experimental Analysis of Behavior*, 22(2):261–273, 1974. 2.2.3

[137] M. Sidman and W. Tailby. Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of behavior*, 37(1):5–22, 1982. 2.2.3

[138] A. Skrondal and S. Rabe-Hesketh. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press, 2004. 3.1, 3.2.1

[139] D. Sloan and C. Norrgran. A neuroscience perspective on learning. *Chemical Engineering Education*, 50(1):29–37, 2016. 1.1, 2.1.1

[140] G. Smith. *Essential statistics, regression, and econometrics*. Academic press, 2015. 3.1

[141] S. Smith. Linking cognition to brain connectivity. *Nature neuroscience*, 19(1):7–9, 2016. 2.1.3

[142] H. Sompolinsky and O. White. Course 8 - Theory of Large Recurrent Networks: From Spikes to Behavior. In C. Chow, B. Gutkin, D. Hansel, C. Meunier, and J. Dalibard, editors, *Methods and Models in Neurophysics*, volume 80 of *Les Houches*, pages 267–340. Elsevier, 2005. 2.1.1

[143] W. Sossin, J. Lacaille, V. Castellucci, and S. Belleville. Associative learning signals in the brain. *Essence of Memory*, page 305, 2008. 2.1.1

[144] O. Sporns. *Networks of the Brain*. MIT press, 2010. 1.4

[145] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018. 2.2.3

[146] E. Suzuki. Time changes, so do people. *Social science & medicine*, 75(3):452–456, 2012. 3.1

[147] D. Svozil, V. Kvasnicka, and J. Pospichal. Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, 39(1):43–62, 1997. 2.2.1

[148] A. Trefler, N. Sadeghi, A. G. Thomas, C. Pierpaoli, C. I. Baker, and C. Thomas. Impact of time-of-day on brain morphometric measures derived from T1-weighted magnetic resonance imaging. *NeuroImage*, 133:41–52, 2016. 2.1.3, 4.1

[149] E. Tulving. Episodic memory: From mind to brain. *Annual review of psychology*, 53(1):1–25, 2002. 2.1.1

[150] J. W. Twisk. *Applied longitudinal data analysis for epidemiology: a practical guide*. cambridge university press, 2013. 3.1

[151] D. o. E. United Nations and S. Affairs. World population ageing 2017: highlights, 2017. 2.1.2

[152] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al. The WU-Minn human connectome project: an overview. *NeuroImage*, 80:62–79, 2013. 2.1.3

[153] B. Vandeginste, D. Massart, L. Buydens, S. D. Jong, P. Lewi, and J. Smeyers-Verbeke. Chapter 44 - Artificial Neural Networks. In B. Vandeginste, D. Massart, L. Buydens, S. De Jong, P. Lewi, and J. Smeyers-Verbeke, editors, *Handbook of Chemometrics and Qualimetrics: Part B*, volume 20 of *Data Handling in Science and Technology*, pages 649–699. Elsevier, 1998. 2.2.1

[154] K. B. Walhovd, A. M. Fjell, and T. Espeseth. Cognitive decline and brain pathology in aging–need for a dimensional, lifespan and systems vulnerability view. *Scandinavian journal of psychology*, 55(3):244–254, 2014. 4.1

[155] B. Wang, A. Pourshafeie, M. Zitnik, J. Zhu, C. D. Bustamante, S. Batzoglou, and J. Leskovec. Network enhancement as a general method to denoise weighted biological networks. *Nature communications*, 9(1):1–8, 2018. 2.6, 2.2.4, 5.5

[156] S.-C. Wang. Artificial neural network. In *Interdisciplinary computing in java programming*, pages 81–100. Springer, 2003. 2.2.1

[157] M. L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. 4.2

[158] B. T. West, K. B. Welch, and A. T. Galecki. *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC, 2014. 3.2.1

[159] M. Wiering and M. Van Otterlo. Reinforcement learning. *Adaptation, learning, and optimization*, 12(3), 2012. 2.2.3

[160] J.-A. Witt, R. Coras, A. J. Becker, C. E. Elger, I. Blümcke, and C. Helmstaedter. When does conscious memory become dependent on the hippocampus: The role of memory load and the differential relevance of left hippocampal integrity for short-and long-term aspects of verbal memory performance. *Brain Structure and Function*, 224(4):1599–1607, 2019. 2.1.4

[161] World Health Organization. Dementia. `https://www.who.int/news-room/fact-sheets/detail/dementia`, 2019. Accessed: 2020-09-10. 2.1.2

[162] T. Yagi, M. Kanekiyo, J. Ito, R. Ihara, K. Suzuki, A. Iwata, T. Iwatsubo, K. Aoshima, A. D. N. Initiative, J. A. D. N. Initiative, et al. Identification of prognostic factors to predict cognitive decline of patients with early Alzheimer's disease in the Japanese Alzheimer's Disease Neuroimaging Initiative study. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5:364–373, 2019. 2.1.2

[163] S. L. Zeger and K.-Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130, 1986. 3.1

[164] S. Zhang, S. M. H. Bamakan, Q. Qu, and S. Li. Learning for personalized medicine: a comprehensive review from a deep learning perspective. *IEEE reviews in biomedical engineering*, 12:194–208, 2018. 2.2.1

Part II

ARTICLES

PAPER A

A PREDICTIVE FRAMEWORK BASED ON BRAIN VOLUME TRAJECTORIES ENABLING EARLY DETECTION OF ALZHEIMER’S DISEASE

Mofrad, Samaneh A., Lundervold, Arvid and Lundervold, Alexander S. 1 June, 2021. A predictive framework based on brain volume trajectories enabling early detection of Alzheimer’s disease. *Computerized Medical Imaging and Graphics*. Elsevier.

Computerized Medical Imaging and Graphics 90 (2021) 101910

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

A predictive framework based on brain volume trajectories enabling early detection of Alzheimer's disease

Samaneh Abolpour Mofrad[a,c,*], Arvid Lundervold[b,c], Alexander Selvikvåg Lundervold[a,c], for the Alzheimer's Disease Neuroimaging InitiativeData used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf, for the Australian Imaging Biomarkers and Lifestyle Flagship Study of AgeingData used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle Flagship Study of Ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database. The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.

[a] *Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Postbox 7030, 5020 Bergen, Norway*

[b] *The Neural Networks and Microcircuits Research Group, Department of Biomedicine, University of Bergen, Bergen, Norway*

[c] *The Mohn Medical Imaging and Visualization Centre (MMIV), Department of Radiology, Haukeland University Hospital, Bergen, Norway*

ARTICLE INFO

Keywords:
Mixed effects models
Machine learning
Longitudinal data analysis
Alzheimer's disease
Mild cognitive impairment
MRI

ABSTRACT

We present a framework for constructing predictive models of cognitive decline from longitudinal MRI examinations, based on mixed effects models and machine learning. We apply the framework to detect conversion from cognitively normal (CN) to mild cognitive impairment (MCI) and from MCI to Alzheimer's disease (AD), using a large collection of subjects sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Australian Imaging, Biomarkers and Lifestyle Flagship Study of Aging (AIBL). We extract subcortical segmentation and cortical parcellation from corresponding T1-weighted images using FreeSurfer v.6.0, select bilateral 3D regions of interest relevant to neurodegeneration/dementia, and fit their longitudinal volume trajectories using linear mixed effects models. Features describing these model-based trajectories are then used to train an ensemble of machine learning classifiers to distinguish stable CN from converters to MCI, and stable MCI from converters to AD. On separate test sets the models achieved an average of accuracy/precision/recall score of 69/73/60% for converted to MCI and 75/74/77% for converted to AD, illustrating the framework's ability to extract predictive imaging-based biomarkers from routine T1-weighted MRI acquisitions.

1. Introduction

About 50 million people world-wide suffer from dementia (World Health Organization, 2019), with a new case appearing every 3.2 seconds (Prince, 2015). The total cost of dementia care has risen to above one trillion US dollars after 2018 (World Health Organization, 2019; Prince, 2015). The most common form of dementia is Alzheimer's disease (AD), responsible for up to 60 − 70% of cases (Prince, 2015). AD

* Corresponding author at: Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Postbox 7030, 5020 Bergen, Norway.
E-mail address: sam@hvl.no (S.A. Mofrad).

https://doi.org/10.1016/j.compmedimag.2021.101910
Received 17 September 2020; Received in revised form 12 February 2021; Accepted 26 March 2021
Available online 2 April 2021

is aging-related, mostly inflicting people above 60 years. This is a steadily growing age group: in the 20th and 21st centuries, both the overall population levels and life expectancy increased drastically, a trend that shows no sign of stopping. In 2018, for the first time recorded in history, people aged 65 and older outnumbered children five years or younger. Currently, about one in 11 people in the world are above 65, and this is expected to increase to one in six by 2050. The number of people above 80 years is projected to rise from 143 million in 2019 to 426 million in 2050 (United Nations Department of Economic and Social Affairs Population Division, 2017).

There has been extensive research into biological and neurological alterations with aging. It is well-known that aging causes a decline in processing speed, working memory and inhibitory function, as well as atrophy in several brain structures (Park and Reuter-Lorenz, 2009; Brookmeyer et al., 2007). These normally-appearing damages intensify with aging-related diseases, making the discrimination between normal and disease-related aging both challenging and important (Reuter--Lorenz and Lustig, 2005).

Alzheimer's disease is a chronic neurodegenerative disease causing the death of neurons. As neurons commonly do not reproduce or get replaced, preventing damage in the first place is crucial to slow its progression. There is no cure for AD – even moderate forms refuse treatment – but medication can affect patients with mild forms of the disease (Dodel et al., 2013; Montgomery et al., 2003). For this reason, as well as optimizing treatment plans, early disease detection and prediction is crucial (Siemers et al., 2016; Guerrero et al., 2016).

In aging, brain atrophy is normal. However, in dementia certain regions of the brain have increased speed of atrophy (Park and Reuter-Lorenz, 2009; Leong et al., 2017; Rodrigue and Raz, 2004; Lundervold et al., 2019; Chandra et al., 2019). While the distinction between the neurodegenerative changes by normal aging and those that characterise AD is not evident, studies have shown that greater shrinkage in specific brain regions is linked to AD (Leong et al., 2017; Raz, 2000; West et al., 1994). For example, hippocampal volume reductions and ventricular expansions show different patterns in healthy aging and in dementia (Thompson et al., 2004), and both can be considered as imaging biomarkers to investigate the rate of brain deterioration (Leong et al., 2017; West et al., 1994; Raz, 2000). The change in the brain has been quantified with different methods and techniques, such as counting neuronal cell loss in brain regions (West et al., 1994) and by calculating the changes in the volume of the brain regions from neuroimaging data (Leong et al., 2017; Raz, 2000).

Such imaging findings, and the uncertainty in the clinical diagnosis of AD, leads to both a need and a potential for further quantitative and indicative imaging biomarkers. In recent years, researchers have constructed a variety of analysis tools and approaches to investigate the aging process in the brain using MRI data, often including machine learning methods (Falahati et al., 2014; Guerrero et al., 2016; Jack et al., 2008; Klöppel et al., 2008; Scheltens et al., 1992; Shi et al., 2009).

While there have been many promising results, there are several limitations in these methods and approaches. For example, an assumption underlying many of the proposed machine learning approaches is that the data instances in follow-up MRI examinations are independent and identically distributed. However, in longitudinal data there are certainly correlations (Falahati et al., 2014; Ngufor et al., 2019; Lei et al., 2017), and using proper longitudinal analysis designs have some important advantages, such as reducing the confounding effect of between-subject variability and making it possible to use non-independent data. Some recent works have taken this into account (Ngufor et al., 2019; Lei et al., 2017; Huang et al., 2016; Zhang et al., 2012; Lim and van der Schaar, 2018), but additional limitations remain. One limitation that the present study aims to overcome is an assumption underlying many other approaches: that all subjects have the same number of measurements, and, even, that the measurements are recorded over the same time interval lengths for the entire sample set. In practice, these assumptions are often invalid, leading other studies to remove instances from their data set (Zhang et al., 2012).

In the present study, we propose a pipeline that is better adapted to such situations. It is a framework based on a combination of mixed effects models (LME) and an ensemble machine learning model (Fig. 2). We used linear time-dependent mixed effects model parameters to derive representative features from the MRI measurements in the predictive machine learning models. Our approach applies to situations where subjects have varying number of MRI examinations, potentially recorded at different scan intervals. It is also possible to include subjects that were examined at a single time-point. The mixed effects modelling is applied to the volumetry of brain regions computed by FreeSurfer v.6.0 (Fischl, 2012), enabling extraction of subject- and region-specific longitudinal volume trajectories (Fig. 1). The instability and fluctuations observed when analysing brain structure volumes from MRI over time, caused by e.g. computational instabilities, noise, hydration status, scanner upgrades, time-of-day at scanning (Trefler et al., 2016) or slight variation in the acquisition protocol, and not changes in the brain parenchyma per se, become less influential by using a LME model (Fig. 1b). This makes the representation of individual volume trajectories more robust, and the prediction of longitudinal group differences more precise (Bernal-Rusiel et al., 2013).

Our results show the ability of this framework to make early prediction of AD, prior to clinical diagnosis, and, to a certain extent, distinguish between cognitively normal (CN) subjects and those who are at risk of MCI. Such a model-based predictive framework, together with assessments of risk factors, could have great potential in monitoring natural progression and to evaluate effect of possible therapeutic interventions. It can also help the clinician in prognostics and advice regarding lifestyle changes and preparing patients for likely life events of neurodegenerative disease. In a related work by the authors (Mofrad et al., 2021) we have demonstrated the proposed framework's ability to incorporate any kind of longitudinal measure, in that case cognitive measures from psychometric testing, and also that the MRI-derived measures provide additional information to the predictive model.

2. Methods

We applied mixed effects models to derive features from longitudinal MRI examination, and used the features in machine learning models aiming at predicting MCI and AD prior to the clinical events. Our approach has two key parts: (i) feature selection, model development and validation, and (ii) model-evaluation. We used data from ADNI for the first part, and a combination of ADNI and AIBL data for the second, making sure no subjects were used for both model training and evaluation of predictive performance. The use of the AIBL data for evaluation ensured that our models were evaluated on an independent data set, sourced from different institutions and subjects than those represented in the training set. This is a crucial part of evaluating predictive models as one can otherwise easily overestimate such models' generalization abilities. Fig. 2 illustrates our framework, further explained in this section.

2.1. Data

Data used in the preparation of this work were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by principal investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD (Gavidia-Bovadilla et al., 2017). We also used data collected by the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) database (https://aibl.csiro.au). Launched in 2006, AIBL is the largest study in Australia to discover biomarkers, cognitive characteristics, health and lifestyle factors

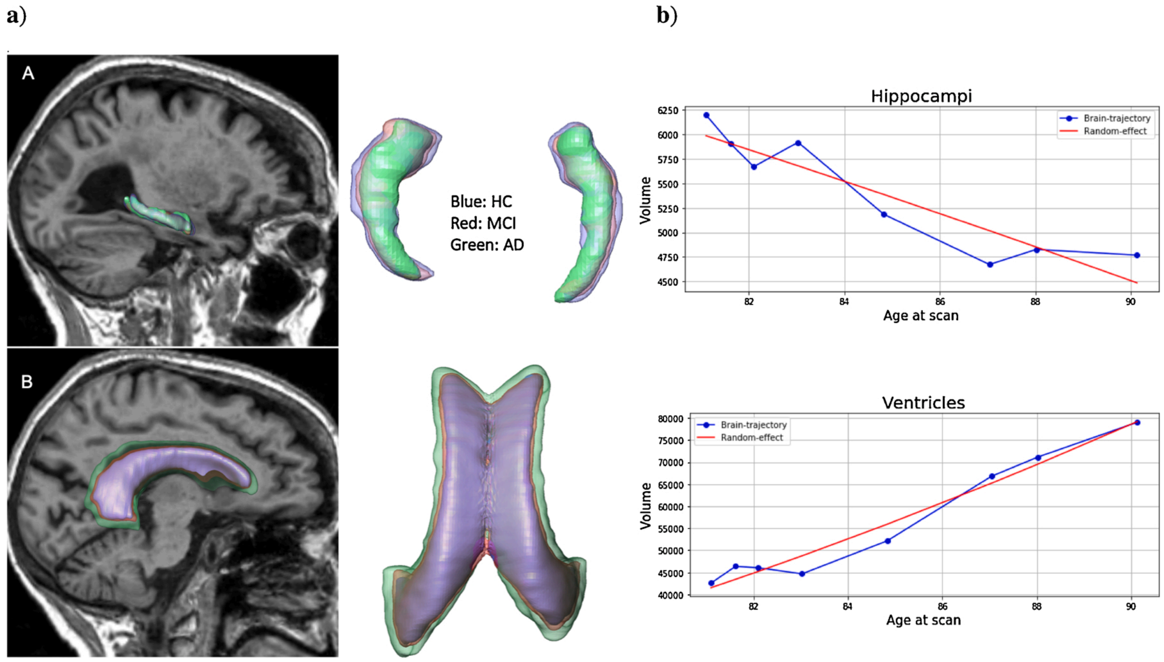


Fig. 1. Aging causes morphometric changes in the brain and dementia accelerate these changes. (**a**) Here we illustrate volume reduction of the hippocampi: left + right hippocampus (A) and expansion of the lateral ventricles (B) with surface renderings from three scans in the series of eight examinations of the same subject. (**b**) A LME model was used to derive representations (i.e. random effects) of such volume trajectories. The blue lines are observed volume trajectories and the red lines are the estimated random effects, based on the eight measurements. Note the small fluctuations or instabilities in the measurements connected by the blue line segments. See the Methods section for more details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

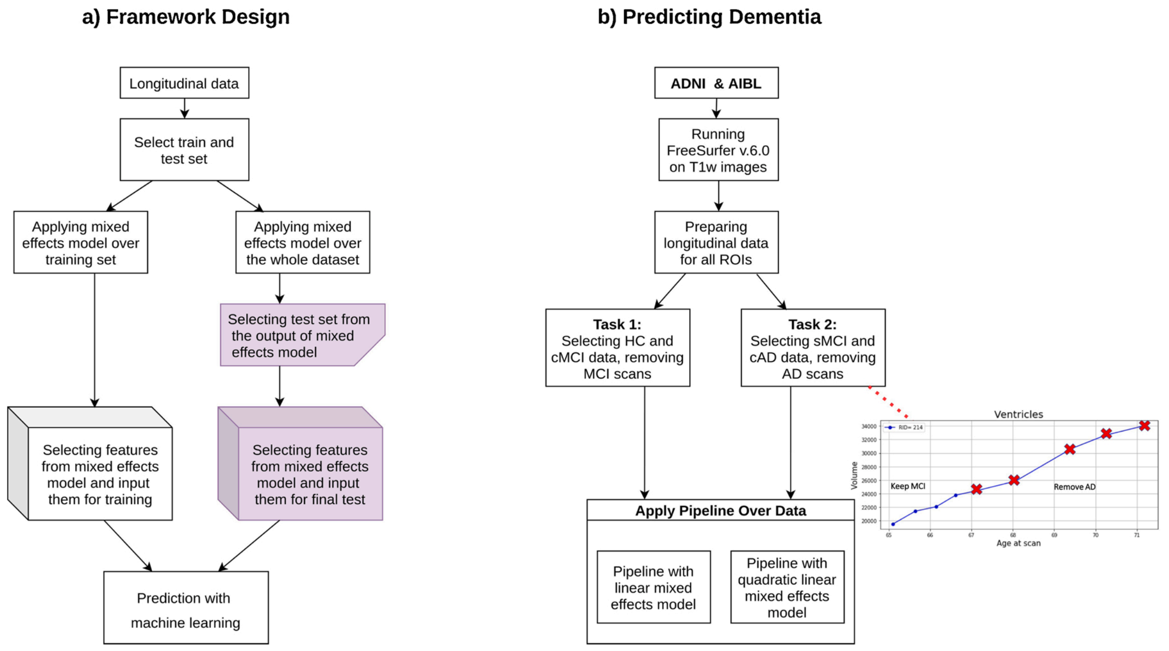

a) Framework Design
Longitudinal data
Select train and test set
Applying mixed effects model over training set
Applying mixed effects model over the whole dataset
Selecting test set from the output of mixed effects model
Selecting features from mixed effects model and input them for training
Selecting features from mixed effects model and input them for final test
Prediction with machine learning
b) Predicting Dementia
ADNI & AIBL
Running FreeSurfer v.6.0 on T1w images
Preparing longitudinal data for all ROIs
Task 1: Selecting HC and cMCI data, removing MCI scans
Task 2: Selecting sMCI and cAD data, removing AD scans
Apply Pipeline Over Data
Pipeline with linear mixed effects model
Pipeline with quadratic linear mixed effects model
Ventricles
RID= 214
Volume
34000
32000
30000
28000
26000
24000
22000
20000
Keep MCI
Remove AD
65
66
67
68
69
70
71
Age at scan

Fig. 2. **a**) Predictive framework for longitudinal data: we first put aside a test set from the longitudinal data. Each mixed effects model is applied separately on the entire data set and on the training set. The features associated with the test set were computed based on constructing the mixed effects models from the entire data set. We used cross-validation on the training set for machine learning model selection. **b**) Prediction of dementia: we first ran FreeSurfer v.6.0 on the longitudinal data from ADNI and AIBL. Then we prepared a table of volumes for brain regions and other information of the subjects. For detecting MCI in task 1, described in Section 3.1, we selected HC and cMCI subjects and removed scans labelled as MCI for all subjects. For detecting AD in task 2, described in Section 3.2, we selected sMCI and cAD subjects and removed scans labelled AD. Finally, a pipeline based on linear time-dependent mixed models was applied.

determining the development of symptomatic AD. It comprises more than 1000 participants with a minimum age of 60 years and contains healthy volunteers, MCI and AD subjects. AIBL study methodology has been reported previously by Ellis et al. (Ellis et al., 2009).

From these cohorts we used longitudinal brain MRI data from subjects scanned multiple times (at least twice) over a period of 15 years. Our data collection consisted of 1673 subjects (with a total of 8002 scans) from ADNI (7764 scans from 1603 subjects) and AIBL (238 scans from 70 subjects).

2.2. Volumetric biomarkers

The ADNI data release contains derived subcortical and cortical measures computed using FreeSurfer on the T1-weighted MR images. FreeSurfer is a powerful, widely used software package providing automated analyses of structural and functional neuroimaging data from cross-sectional or longitudinal studies (Fischl, 2012).

However, the FreeSurfer data released by ADNI is based on two different software versions, v.4.3 (from March 2009) and v.5.1 (May 2011), both of which are superseded by v.6.0 released in January 2017. Previous studies have demonstrated significant discrepancies between different versions of FreeSurfer (Chepkoech et al., 2016; Gronenschild et al., 2012; Klauschen et al., 2009), and our own exploratory data analyses based on the ADNI data also demonstrate such an effect. For example, Fig. 3a indicates the dissimilarity of volume measurement with the two versions of FreeSurfer the ADNI consortium used on their data, v.4.3 and v.5.1. The results in Fig. 3 show clear discrepancies between hippocampus volumes derived from scanners of field strengths 1.5 Tesla and 3.0 Tesla. This highlights the importance of not changing the version of FreeSurfer during longitudinal studies, especially those involving scanners of different field strengths.

To get more precise information about the potential negative impact of the varying FreeSurfer versions, we conducted an experiment using FreeSurfer v.5.3 and v.6.0 (the newest version at the time of experiment). We selected 80 subjects, controlling for disease status (CN/Dementia, 40/40), gender (F/M, 40/40), scanner field strength (1.5T/3T, 40/40), and age ([75,80)/[80,85), 40/40). One of the results is shown in Fig. 4, indicates that the effect of FreeSurfer versions differs between CN and Dementia subjects. We concluded that the FreeSurfer version is an important factor for studying atrophy, especially in the small regions of the brain (e.g. the hippocampus), and therefore reprocessed all the ADNI and AIBL data using FreeSurfer v.6.0 on Ubuntu 18.04 GNU/Linux workstations. This gave us the data set used in the remainder of this work.

2.3. Mixed effects models

Our framework is based on linear mixed effects models (LME), a well-established approach to longitudinal data analysis, used to derive regression models from dependent data. In contrast to simpler linear models, LME provides a combination of fixed and random effects as predictor variables (Bell and Jones, 2015; Harrison et al., 2018; Lindstrom and Bates, 1990; Muller et al., 2013; West et al., 2014). Mixed effects models allow the collection of relatively simple, robust, noise-free, and subject-specific representations of brain change over time, as illustrated by the red lines in Fig. 1b, based on age at scan as the covariate.

As some brain ROI volumes versus time show linear cohort behavior while others behave nonlinearly (cf. Fig. 5), we were motivated to use LME models with both linear and nonlinear (quadratic) covariates. Our models are based on the model presented by West et al. (West et al., 2014), also used in our previous work (Lundervold et al., 2019):

$$\mathrm{Vol}^r_{\mathrm{ij}} = \underbrace{\beta^r_0 + \beta^r_1 \mathrm{Age}_{\mathrm{ij}}}_{\mathrm{fixedeffect}} + \underbrace{b^r_{0i} + b^r_{1i}\mathrm{Age}_{\mathrm{ij}} + \varepsilon^r_{\mathrm{ij}}}_{\mathrm{randomeffect}}, \quad (1)$$

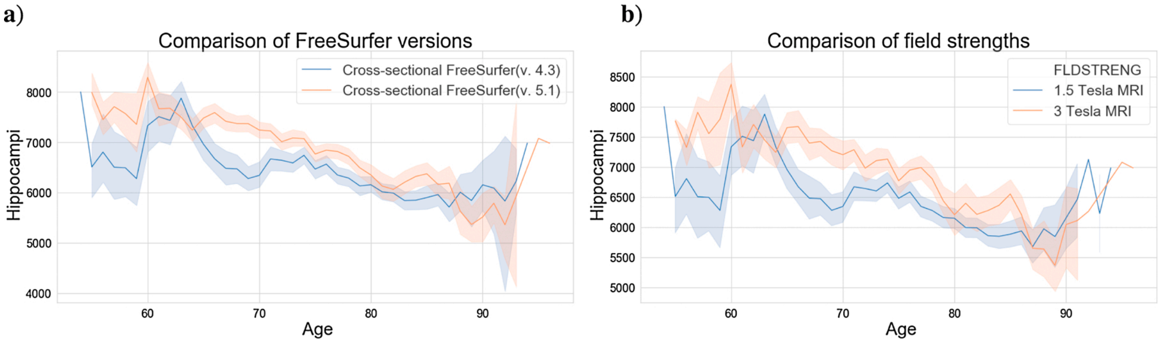


Fig. 3. Plotting the hippocampi volumes for all ADNI subjects across age indicates a discrepancy between (**a**) the volumes calculated by different versions of FreeSurfer and (**b**) the volumes recorded from MRI scanners having different field strengths.

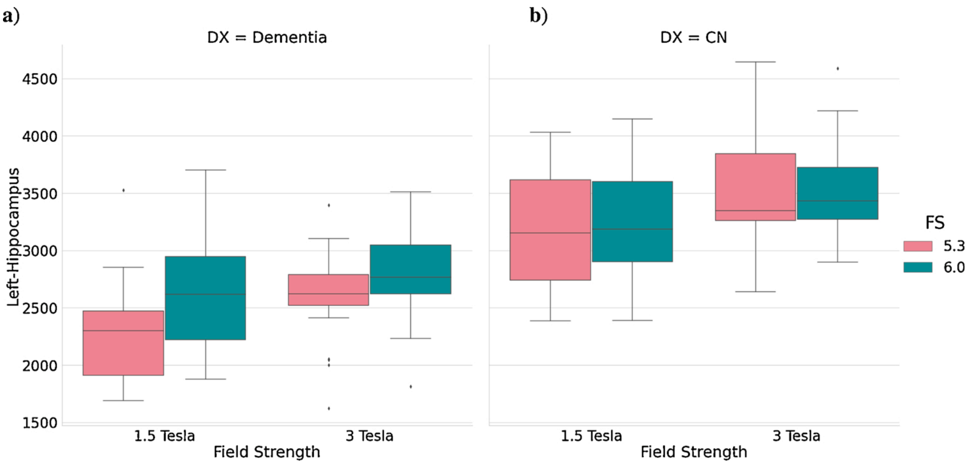


Fig. 4. Box-plots illustrating the importance of FreeSurfer version and magnetic field strength on measuring the volume of the left hippocampus. Each paired box-plot, blue and yellow, contains the same T1w volumes processed with FreeSurfer v. 6.0 and v. 5.3, respectively. **a)** shows volume difference for subjects diagnosed with dementia. **b)** shows volume difference for CN subjects. For dementia the volume discrepancies between FreeSurfer versions are both large and statistically significant (paired t-test, p <0.05) for both 1.5 and 3 Tesla scanners. For CN the version-related differences are insignificant. Note that while we have controlled the gender and age in these groups (1.5 and 3 Tesla) the subjects are different, which makes a precise conclusion of the impact of varying scaner versions difficult. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

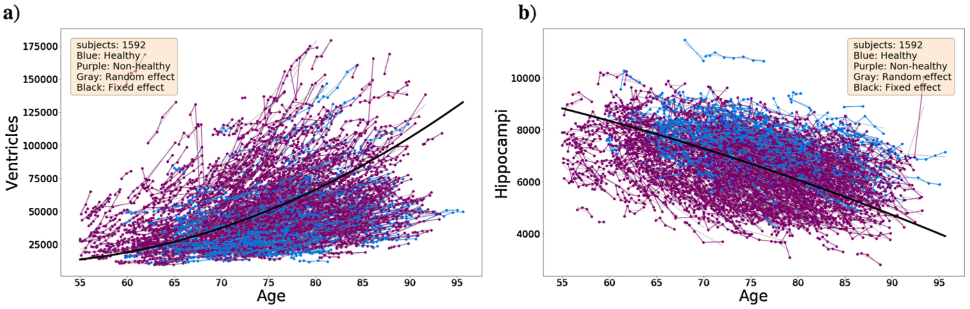


Fig. 5. Longitudinal trajectories for the eTIV-normalised volumes of ventricles (**a**) and hippocampi (**b**) versus age at scan. Healthy subjects are marked in blue, non-healthy in purple. The thick black curve is the cohort fixed effect regression line. The random effects computed by Eq. (2) are shown as thin grey lines for each subject. Note the steady decrease in the hippocampal volumes with time in the age range 55–95 years, and the concomitant nonlinear increase in the lateral ventricle volumes. The plot indicates that the most extensive tissue losses are found among subjects labelled as not cognitively normal. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

where r denotes the brain region, $\mathrm{Vol}^r_{\mathrm{ij}}$ is the volume of region r for subject $i: 1,\ldots,N$ at scan $j: 1,\ldots,n_i$. In our case, n_i varies between 2 and 11. $\mathrm{Age_{ij}}$ is age (in years) of subject i at scan j. This is the only predictor variable in the model. The β^r_0 and β^r_1 are fixed effect parameters, while b^r_{0i} and b^r_{1i} are random effects parameters and $\varepsilon^r_{\mathrm{ij}}$ denotes the random residual errors.

As seen in Fig. 5, the cohort volume change in the lateral ventricles demonstrate a quadratic behavior, likely due to atrophy over time in multiple brain regions leading to an enlargement of the cerebrospinal fluid-filled lateral ventricles, compensating the tissue loss (i.e. total intracranial volume is preserved). To model this behaviour, we assume the rates of volume change are covariant with both `age` and `age`^{2}. Accordingly, our mixed effect models are:

$$\mathrm{Vol}^r_{\mathrm{ij}} = \underbrace{\beta^r_0 + \beta^r_1\mathrm{Age_{ij}} + \beta^r_2\mathrm{Age}^2_{\mathrm{ij}}}_{\mathrm{fixedeffect}} + \underbrace{b^r_{0i} + b^r_{1i}\mathrm{Age_{ij}} + b^r_{2i}\mathrm{Age}^2_{\mathrm{ij}} + \varepsilon^r_{\mathrm{ij}}}_{\mathrm{randomeffect}}, \tag{2}$$

where $(\beta^r_0, \beta^r_1, \beta^r_2)$ are fixed effect parameters and $(b^r_{0i}, b^r_{1i}, b^r_{2i})$ are random effect parameters.

We used the `mixedlm` function in the Python `statsmodels` library (Seabold and Perktold, 2010) (version 0.11.0) to construct and fit the LME models to the data, extracting a mean cohort trajectory (fixed effect) and the subject-specific trajectories (random effects). In this setting the model is linear in the parameters, but can be nonlinear in the covariates. The model parameters (βs and bs) were estimated and stored for each subject, according to Eq. (1) and Eq. (2).

Fig. 5 shows fixed and random effects regressions computed by Eq. (2) for subjects split into two groups: *healthy* (HC, n = 407, f/m = 215/192) and *non-healthy* (sMCI, cAD and sAD, n = 1185, f/m = 492/693). It shows a difference between normal age-related atrophy (blue) and increased atrophy in the case of neurodegenerative disease (purple). This figure indicates the potential of our approach of deriving features from LME models for classifying our different subgroups defined in Table 1a.

We used the volume increase of the ventricles as a measure of total brain atrophy and the volume change in the hippocampi, as it is a well-known structure affected by dementia.

Derived features

From the mixed effects models we derived four features for each individual ROI trajectory: (i) For the linear models (Eq. (1)), a vector of random effect covariates, (b^r_{0i}, b^r_{1i}), containing the intercept of the group and the slope of the random effects line. For the nonlinear models (Eq. (2)) we used the vector $(b^r_{0i}, b^r_{1i}, b^r_{2i})$, the intercept for the group and the coefficients of `age` and `age`^{2}; (ii) and (iii) The deviation measured at the first scan, d^0_i, and at the last scan, $d^{n_i}_i$, respectively. In other words, the derived random effects values at the first and the last scans, (illustrated in Fig. 6), as given by

Table 1
a) The original ADNI class labels and the longitudinal labels used in the present study with their descriptions. **b)** Total number of subjects and number of T1-weighted MR images according to class label in our study, selected from ADNI and AIBL.

a) Source	Class	Class description
ADNI	CN	Cognitively normal at visit
	MCI	Mild cognitive impairment at visit
	Dementia	Alzheimer's disease at visit
Our study	HC	CN at all visits
	cMCI	Initially CN, later converted to MCI
	rHC	Risky CN: cMCI with MCI scans removed
	sMCI	MCI at all visits
	cAD	Initially MCI, later converted to Dementia
	rMCI	Risky MCI: cAD with Dementia scans removed
	sAD	Dementia at all visits

b) Class	ADNI		AIBL	
	ID	#Images	ID	#Images
HC	407	1994	24	90
cMCI	109	596	24	80
sMCI	509	2500	11	34
cAD	269	1540	11	34
sAD	298	1055	-	-
ALL	1603	7764	70	238

$$\mathrm{d}^{\mathrm{j}}_{\mathrm{i}} = \mathrm{Vol_{ij}} - (\beta_0 + \beta_1\mathrm{Age_{ij}}) \tag{3}$$

and, for the nonlinear models, d^0_i and $d^{n_i}_i$ given by

$$\mathrm{d}^{\mathrm{j}}_{\mathrm{i}} = \mathrm{Vol_{ij}} - (\beta_0 + \beta_1\mathrm{Age_{ij}} + \beta_2\mathrm{Age}^2_{\mathrm{ij}}); \tag{4}$$

where j is either 0 or n_i. (iv) The difference of volumes at the first and last scans, divided by the number of years between them (Eq. (5), i.e. the slope of atrophy from the first to the last measurement):

$$\mathrm{Atrophy}^{\mathrm{slope}}_{\mathrm{i}} = \frac{\mathrm{V}_{\mathrm{in_i}} - \mathrm{V_{i0}}}{\mathrm{Age}_{\mathrm{in}_i} - \mathrm{Age}_{i0}} \tag{5}$$

where $\mathrm{V_{i0}}$ and $\mathrm{V_{in_i}}$ are the volumes at the first and last scans for subject i, respectively. Feature (iv) is motivated by the varying number and timing of scans for the subjects, and that the atrophy seen over e.g. 10 years for one subject can be equal to the atrophy in two years for another (see Fig. 6).

2.4. Predictive models

As input features to our machine learning models we used the subjects' sex, average age at scans, age at last scan, and the above four features from mixed effects models, scaled according to

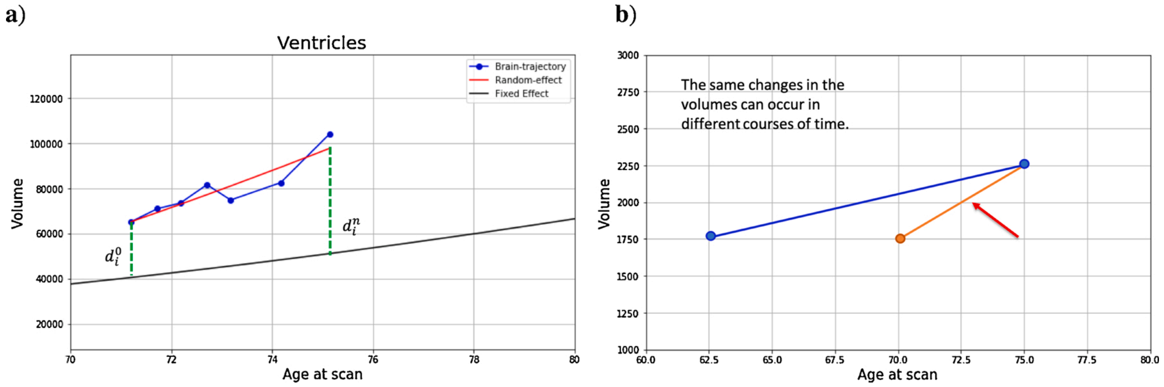


Fig. 6. **a**) Estimated values for random effects at first and last scans (d_i^0, d_i^n) are considered as features (ii) and (iii), respectively. **b**) The linear slop of atrophy (Eq. (5)) calculated based on volumes at first and last scans. The time points are different, and therefore the amount of atrophy in 10 years for a subject can be the same as a 5 years atrophy for another subject (see the red arrow). Therefore, the slope of total atrophy in each ROI is considered as a feature, (iv), for each subject.

$$\text{standardscaling}: \quad \tilde{\mathbf{x}} = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\sigma}, \quad \text{ormax} - \text{minscaling}: \quad \tilde{\mathbf{x}} = \frac{\mathbf{x} - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})},$$

where x is a vector of features, $\bar{\text{x}}$ is the mean value of vector x, and σ is the standard deviation for x. We trained an ensemble of a logistic regression and a support vector machine, based on a soft voting strategy, i.e. using the weighted average probabilities from each model in the ensemble. Rather than using single models, with their own specific decision logic, an ensemble constructed from multiple diverse, individually-tuned models can result in a more robust, higher-performing model (Dietterich, 2000; Saeys et al., 2008). We used recall and accuracy scores to assess our models during development and hyper-parameter selection, using subject-level cross-validation on the training set. For each model we set up a grid search through sets of hyperparameters, attempting to find the models with the best generalization abilities.

For the support vector classification model (SVC) we evaluated the regularization parameter *C*, polynomial, sigmoid and radial basis function kernels, and the kernel coefficient. In `scikit-learn`, the kernel coefficient γ is either set to *scale* or *auto*. For training data with length *n*, the *scale* setting means that the model uses $1/(\#\text{features} \times \text{variance}(\mathbf{x}))$ as the value of γ and *auto* means it uses $1/(\#\text{features})$. For the logistic regression model we evaluated whether to include an l_2 penalty and the strength of this regularization (*C*). For both SVC and logistic regression we fixed a random seed, to ensure reproducibility, and set the maximum number of iterations to 500.

We performed feature selection and model development using T1-weighted images from the ADNI dataset, and model evaluation with data from non-overlapping subjects sourced from both ADNI and AIBL. When constructing predictive models for conversion from healthy to MCI and from MCI to AD, we removed all MRI measures taken from the time of conversion and after.

We considered two predictive tasks, described using the subject classes of Table 1:

1 HC subjects (n = 133, f/m = 56/77) versus converted to MCI subjects (cMCI, n = 133, f/m = 55/78),
2 stable MCI subjects (sMCI, n = 279, f/m = 114/165) versus converted to AD subjects (cAD, n = 279, f/m = 111/168).

In task 1 we removed the MRI scans that corresponded to clinical diagnoses of MCI from the cMCI subjects. We call the resulting collection *risky HC* (rHC). In task 2 we removed MRI scans corresponding to AD from the cAD subjects, calling the resulting collection *risky MCI* (rMCI). Details about diagnosis labels and the number of subjects are given in Table 1.

3. Results

3.1. Task 1: Prediction of MCI

We applied our model to two groups of subjects: the subjects marked as cognitively normal at all visits (HC) and the risky HC (rHC, i.e. MRI data from cMCI subjects obtained by removing the scans clinically labelled as MCI). The goal was to investigate whether regular MRI scans can separate HC from rHC, as early detection of MCI based on brain morphometry is an important but also a very challenging task. The subject trajectories for ventricles and hippocampi (Fig. 7) found using LME model show atrophy in the hippocampi and volume increase in the ventricles during aging, while also showing similarity in the trajectories of HC and cMCI. In addition, Fig. 8 shows similar behavior for the average volume of ventricles and hippocampi in HC and cMCI groups of participants, indicating the difficulty of the classification task.

We used data from ADNI for training and data from AIBL for model evaluation. After optimizing the model based on leave-one-out cross validation over the entire training data set (details of hyper-parameters are shown in Table 2 and also in the accompanying code repository[1]), we performed a 15 fold cross validation experiment on the training data set, controlling for labels, age, and gender in the hold-out folds. The mean accuracy and standard deviation obtained by the 15 folds for ventricles and hippocampi ranged from 69 ± 6% to 73 ± 7% (see Table 3 for more details). We then applied the model on the main test set from AIBL for evaluation.

We evaluated the model with eight different feature vectors. First, we extracted four sets of features from the ventricles and the hippocampi volumes, using linear and quadratic LME models. Then we applied the ensemble model on each set of features to find the ones with the highest classification performance. We obtained the best accuracy (71%) for quadratic features extracted from hippocampi. See Table 3 for details about these results.

Next, we combined the extracted features of ventricles and hippocampi to see whether this would improve the classification. The results are shown in Table 3.

3.2. Task 2: Prediction of AD

The ability to predict AD before the symptoms are caught by the clinician is the main objective for our study. We selected sMCI and cAD subjects from ADNI and AIBL to investigate to what extent the atrophy trajectories can distinguish the stable MCI from the risky MCI (subjects

[1] https://github.com/MSamane/A-predictive-framework-for-Alzheimers-disease

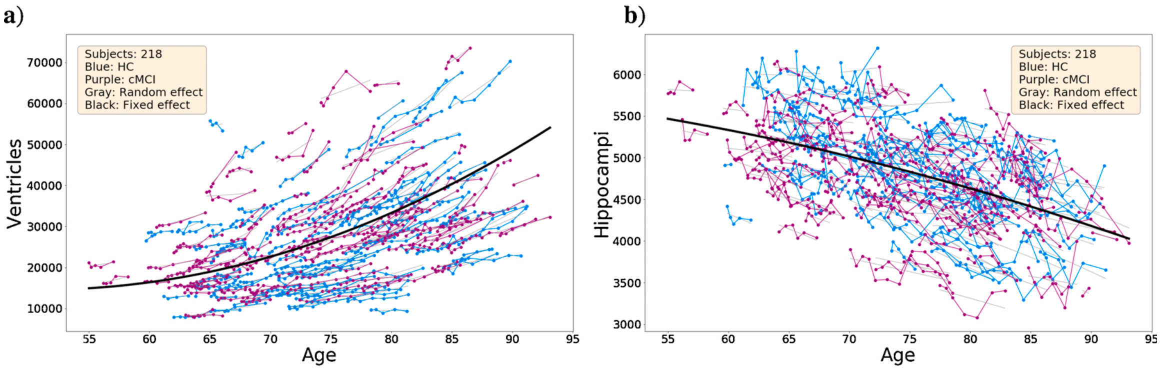


Fig. 7. Task 1: Longitudinal trajectories for the eTIV-normalised volumes of the lateral ventricles (**a**) and hippocampi (**b**) versus age at scan. The thick black curve is the cohort nonlinear regression line. The random effects computed by Eq. (2) are shown as thin grey lines for each subject. The volume of the hippocampi decreases over time, likely contributing to the increase in the lateral ventricle volumes.

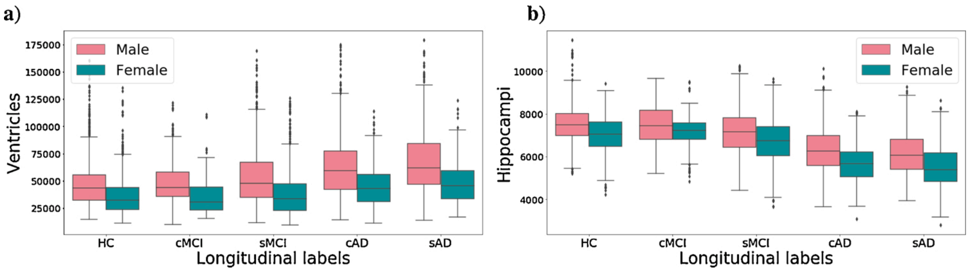


Fig. 8. Ventricles and hippocampi volumes (mm^3) versus our longitudinal subjects from the ADNI dataset, labelled according to Table 1a, show a similarity between HC and cMCI subjects. The sMCI and cAD show a difference in their ventricle volume expansions and their hippocampi atrophy. The difference between ROIs volumes for males and females indicates that gender is an important factor when comparing brain volumes.

Table 2
Model hyperparameters for task 1, for different ROIs-feature combinations, obtained by applying leave-one-out cross validation on the training set.

HC vs. rHC	LME covariates	Logistic regression		SVC		
ROI		scaler	C	scaler	C	kernel
Ventricles	linear	standard	3.13	standard	4.0	poly
	nonlinear	standard	7.78	standard	15.56	poly
Hippocampi	linear	standard	6.7	minmax	7.525	poly
	nonlinear	standard	11.16	standard	10	rbf
Combination	linear	standard	5.6	minmax	6.7	poly
	nonlinear	standard	4.5	minmax	8.9	poly
	nonlin vent, lin hipp	standard	4.5	minmax	20	poly
	lin vent, nonlin hipp	standard	20	minmax	6.7	poly

obtained by removing scans clinically labelled as AD from the cAD subjects). On average, the trajectories for ventricles and hippocampi show a clear atrophy in the hippocampi and a volume increase of the ventricles (Fig. 9). Furthermore, the blue subject trajectories (sMCI) in Fig. 9b are on average above the purple trajectories (cAD).

As for task 1, we first optimized the ensemble machine learning model using leave-one-out cross validation over the entire training data set from ADNI (details of hyper-parameters are in Table 4 and in the accompanying code repository), and then performed a 15 fold cross validation experiment on the training data set, controlling for labels, age, and gender in the hold-out folds, before evaluating the model on a test set. The mean accuracy and standard deviation ranged from 77 ± 4% to 79 ± 6% (see Table 5).

As there are few cAD subjects in AIBL, the test set was constructed using subjects from ADNI (n = 99) and AIBL (n = 22). As for task 1, we applied the ensemble models on eight sets of features extracted by linear

Table 3
Classification results for task 1 for the different ROI features. Note that the accuracy obtained in the 15 fold cross validation experiment is on average better than the accuracy in the final test set sourced from AIBL. As the training and hold-out data in the cross validation are both sourced from ADNI, while the test set is based on AIBL, this is perhaps not surprising.

HC vs. rHC	LME covariates	CrossVal Acc (%)	Accuracy (%)	Precision (%)		Recall (%)		F_1 score (%)	
ROI				HC	rHC	HC	rHC	HC	rHC
Ventricles	linear	69 ± 4	69	65	76	83	54	73	63
	nonlinear	69 ± 6	69	67	71	75	62	71	67
Hippocampi	linear	73 ± 7	67	64	70	75	58	69	64
	nonlinear	71 ± 6	71	67	78	83	58	74	67
Combination	linear	70 ± 7	73	69	79	83	62	75	70
	nonlinear	70 ± 8	65	64	65	67	62	65	64
	nonlin vent, lin hipp	71 ± 8	65	64	65	67	62	65	64
	lin vent, nonlin hipp	72 ± 8	73	69	79	83	62	75	70

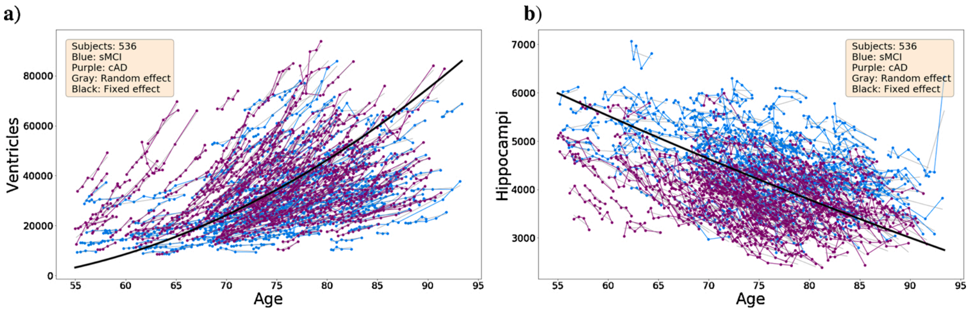


Fig. 9. Task 2: Longitudinal trajectories for the normalised volumes of ventricles (**a**) and hippocampi (**b**) versus age at the scan. The thick black curve is the cohort nonlinear regression line. The random effects computed by Eq. (2) are shown as thin grey lines for each subjects. The volume of hippocampi decreases over time contributing to the increase in ventricular volume. The plot indicates that, the most extensive losses are found among cAD subjects.

Table 4
Model hyper parameters for task 2, for different ROIs-feature combination, obtained by applying leave-one-out cross-validation using the training set.

sMCI vs. rMCI	LME covariates	Logistic regression		SVC		
ROI		scaler	C	scaler	C	kernel
Ventricles	linear	standard	19.9	standard	6.72	rbf
	nonlinear	minmax	8.9	minmax	4.5	poly
Hippocampi	linear	standard	8.89	minmax	11.12	poly
	nonlinear	standard	7.78	minmax	2.23	poly
Combination	linear	standard	6.7	minmax	2.3	poly
	nonlinear	standard	4.5	minmax	5.6	poly
	nonlin vent, lin hipp	standard	7.8	minmax	3.4	poly
	lin vent, nonlin hipp	standard	4.5	minmax	1.2	poly

and quadratic LME from ventricles and hippocampi. The results are presented in Table 5 and in the confusion matrices in Fig. 10 and Fig. 11. The highest accuracy, 78%, was obtained when combining the quadratic features from the hippocampi and ventricles.

4. Discussion

We have developed a flexible and simple framework for extracting features and constructing predictive models from longitudinal MRI in relation to cognitive aging and dementia, based on mixed effects models and ensemble machine learning methods. A strength of the approach is its inherent ability in tackling longitudinal data sets, including situations with sets of subjects with a varying number of scans, taken at different time intervals, which is a common occurrence in longitudinal studies.

We applied the framework to predict dementia, using a large data set sourced from ADNI and AIBL for training and testing. Based on mea-

Table 5
Classification results for task 2, related to different ROI's features. The 15-folds validation results are based on only ADNI dataset (subset of training set) while the other results are based on final test set, a combination of subjects from ADNI and AIBL data.

sMCI vs. rMCI	LME covariates	CrossVal Acc(%)	Accuracy(%)	Precision(%)		Recall(%)		F_1 score(%)	
ROI				sMCI	rMCI	sMCI	rMCI	sMCI	rMCI
Ventricles	linear	77 ± 4	74	77	71	67	80	72	75
	nonlinear	77 ± 4	73	78	69	64	82	70	75
Hippocampi	linear	79± 5	74	78	70	66	82	71	75
	nonlinear	78 ± 5	77	77	77	77	77	77	77
Combination	linear	79 ± 5	75	76	74	74	77	75	75
	nonlinear	79 ± 4.5	78	74	82	85	70	79	76
	nonlin vent, lin hipp	78 ± 5	76	78	75	74	78	76	76
	lin vent, nonlin hipp	79 ± 6	74	73	76	79	70	76	73

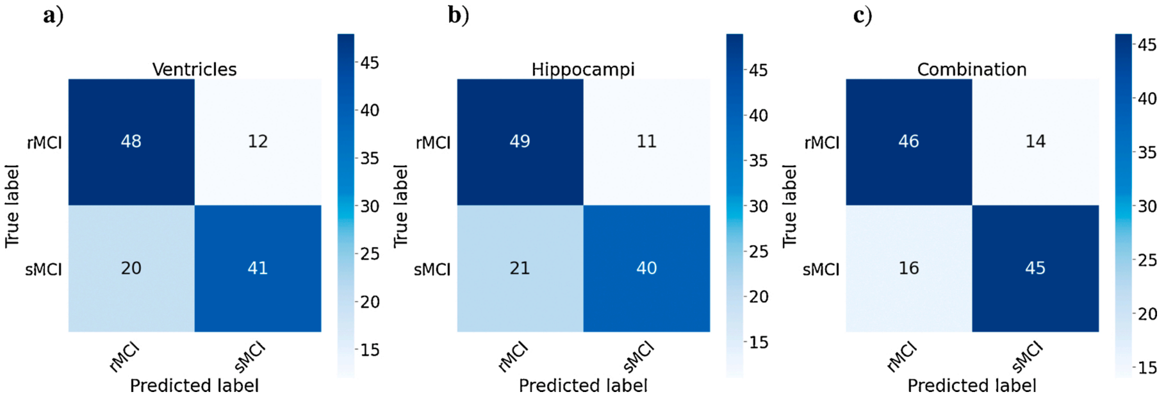


Fig. 10. Confusion matrices for classification of sMCI vs. rMCI based on features extracted from LME model (Eq. (1)) for the ventricles (**a**), the hippocampi (**b**), and the combination of ventricles and hippocampi (**c**).

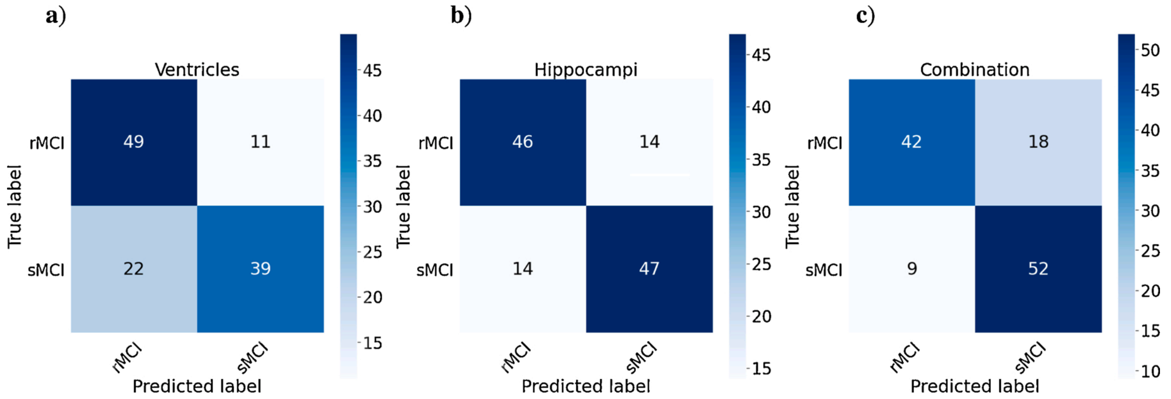


Fig. 11. Confusion matrices for classification of sMCI vs. rMCI based on features extracted from quadratic mixed effects model (Eq. (2)) for the ventricles (**a**), the hippocampi (**b**), and the combination of ventricles and hippocampi (**c**).

surements of hippocampal and lateral ventricle volumes in single subjects over time, we were able to make predictions of conversion from cognitively normal (CN) to mild cognitive impairment (MCI) and from stable MCI to AD, ahead of the corresponding clinical diagnoses, with accuracies of 73% and 78%, respectively. The task of predicting conversion from healthy to MCI is inherently difficult, as it is very challenging to differentiate cognitive decline related to MCI symptoms from cognitive decline with stable cognitive performance at the baseline (Yue et al., 2021). Therefore, our above chance level results at this task is notable. Since the subjects in our study vary with respect to the number of MRI scans and number of years between scans, it is not straightforward to state how early we can predict the risk of MCI or AD prior to diagnosis. In our sample the average time interval between the MRI scans for each subject is 0.53 years (HC: 0.58, cMCI: 0.62, sMCI: 0.48, cAD: 0.53). Therefore, we cannot expect to obtain predictions of conversion to MCI or AD earlier than half a year ahead of the actual conversion.

There are a few studies that predict the conversion from HC to MCI using multi-domain features, including MRI scans (Mofrad et al., 2021; Yue et al., 2021; Albert et al., 2018). Albert et al. (2018) employed imaging-biomarkers related to the hippocampus and the entorhinal cortex in a sample of 224 subjects (178 HC vs. 46 cMCI) obtaining a sensitivity of 64% in predicting the conversion to MCI. Yue et al. (2021) obtained an accuracy/sensitivity of 63%/42% in predicting decline to MCI using MRI-derived features only, improving their results to 70% accuracy and 63% sensitivity when incorporating multi-domain features. Regarding conversion from MCI to AD, Young et al. (2013) predicted this conversion within three years with a 74% accuracy using a Gaussian process classification. This is on par with our results of 78% accuracy. Interestingly, using a deep learning approach (CNN and RNN) and longitudinal MRI data Cui et al. (2019) obtained 72% classification accuracy and 76% sensitivity in their experiments to predict pMCI vs. sMCI.

There are several limitations related to the available data material in our study and in our methods. For example, the group of patients with MCI is highly diverse (Cole and Franke, 2017; Walhovd et al., 2014; Nyberg and Pudas, 2019), and a clinical diagnosis of Alzheimer's disease is inherently uncertain, as the disease is only definite post-mortem (Association, 2013). This is not captured by the labels in ADNI and AIBL, and also holds for similar studies mentioned above.

Furthermore, variability of non-biological origin in MRI measurements, occuring between subjects and in subject examinations over time, will take place (different scanners, calibration issues and scanner drift, different head positions, head motion during scan, etc.) (Trefler et al., 2016; Di et al., 2019). There are also instabilities and uncertainties in the algorithms, libraries and numerical schemes used to compute brain region-specific measures that will lead to sources of variation affecting predictive models and their performance. In this context, we used FreeSurfer v.6.0 and v.5.3 to compute the volumes of the hippocampi and lateral ventricles, exploring some of the inherent variation when using different version of the software and when the images are recorded on scanners of different magnetic field strength (Fig. 3 and Fig. 4). Based on this exploration, we re-computed the volumes in the ADNI and AIBL data sets using the same version of FreeSurfer (v.6.0) to reduce some of the variability. But some instability issues surely remain.

In this work we focused on establishing a framework using only MRI-based morphometric measurements of the hippocampi, as a brain region well-known to be impacted by dementia (Leong et al., 2017; Chandra et al., 2019; Rodrigue and Raz, 2004; Raz, 2000), and the lateral ventricles, as a global measure (proxy) of brain atrophy (Leong et al., 2017; Chandra et al., 2019). Other regions are also impacted by aging and dementia, and inclusions of measures from those ROIs could potentially lead to improved predictions (Rodrigue and Raz, 2004; Raz, 2000; Leong et al., 2017; Hensel et al., 2005; Poulin et al., 2011).

Another approach taken by some researchers (Cui et al., 2019) is to train convolutional and recurrent neural networks to make predictions directly from subjects' MRI recordings (see e.g. Wen et al., 2020, for an overview). This has the possible advantage of bypassing a lot of careful feature engineering and feature selection with its inherent issues, while still making as accurate or more accurate predictions. But it suffers from the disadvantage of leading to less explainable models (Lundervold and Lundervold, 2019).

A major opportunity and motivation for applying machine learning to neuroimaging examinations in middle aged or elderly subjects that are at risk of cognitive decline, mild cognitive impairment or full blown AD, is the ability to make predictions for single individuals. Such imaging procedures and data analysis will thus support *personalized medicine*, and with detailed quantification of image-derived features in combination with subject-specific information obtained from other sources, one can also aim for *precision medicine*. A contribution of the present work is the design and testing of an expressive and flexible machine learning framework that supports both longitudinal image-derived features as well as cognitive scores (Mofrad et al., 2021), where biochemical measures, genetic profiles and other clinical or laboratory measurements can be included. In the context of the present work and available data in the used data repositories, further improvements could potentially be made by including features from multi-modal MRI, such as functional BOLD MRI (Sperling, 2011; Lajoie et al., 2017) and diffusion MRI (Doan et al., 2017), or the presence of the APOE4 gene variant (Kim et al., 2009; Safieh et al., 2019), or values from CSF analyses (Janelidze et al., 2020). Including results from clinical examinations would also be valuable (Holleran et al., 2020), as the present authors have reported in (Mofrad et al., 2021). Challenges for clinical use include the trade-off between locally available measurement techniques and infrastructure (e.g. scanners and protocols), the need for feasible patient examination times, the quality and management of model predictions in single individuals, and the consideration of available options for therapy and interventions.

Declaration of interests

None.

Authors' contribution

A.S.L and A.L. conceived the approach, S.A.M. and A.S.L. conceived the experiments, S.A.M. conducted the experiments and analysed the results. All authors reviewed the manuscript.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

Our work was supported by the Trond Mohn Research Foundation, grant number BFS2018TMT07. We thank Hauke Bartsch for several useful comments and inputs throughout our work, and for assisting with Fig. 1.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

World Health Organization, 2019. Dementia (accessed 10.09.20). https://www.who.int/news-room/fact-sheets/detail/dementia.

Prince, M.J., 2015. World Alzheimer Report 2015: The Global Impact of Dementia: An Analysis of Prevalence, Incidence, Cost and Trends. Alzheimer's Disease International.

United Nations Department of Economic and Social Affairs Population Division, 2017. World Population Ageing.

Park, D.C., Reuter-Lorenz, P., 2009. The adaptive brain: aging and neurocognitive scaffolding. Annu. Rev. Psychol. 60, 173–196.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M., 2007. Forecasting the global burden of Alzheimer's disease. Alzheimer's Dementia 3 (3), 186–191.

Reuter-Lorenz, P.A., Lustig, C., 2005. Brain aging: reorganizing discoveries about the aging mind. Curr. Opin. Neurobiol. 15 (2), 245–251.

Dodel, R., Rominger, A., Bartenstein, P., Barkhof, F., Blennow, K., Förster, S., Winter, Y., Bach, J.-P., Popp, J., Alferink, J., et al., 2013. Intravenous immunoglobulin for treatment of mild-to-moderate Alzheimer's disease: a phase 2, randomised, double-blind, placebo-controlled, dose-finding trial. Lancet Neurol. 12 (3), 233–243.

Montgomery, S.A., Thal, L., Amrein, R., 2003. Meta-analysis of double blind randomized controlled clinical trials of acetyl-L-carnitine versus placebo in the treatment of mild cognitive impairment and mild Alzheimer's disease. Int. Clin. Psychopharmacol. 18 (2), 61–71.

Siemers, E.R., Sundell, K.L., Carlson, C., Case, M., Sethuraman, G., Liu-Seifert, H., Dowsett, S.A., Pontecorvo, M.J., Dean, R.A., Demattos, R., 2016. Phase 3 solanezumab trials: secondary outcomes in mild Alzheimer's disease patients. Alzheimer's Dementia 12 (2), 110–120.

Guerrero, R., Schmidt-Richberg, A., Ledig, C., Tong, T., Wolz, R., Rueckert, D., et al., 2016. Instantiated mixed effects modeling of Alzheimer's disease markers. NeuroImage 142, 113–125.

Leong, R.L., Lo, J.C., Sim, S.K., Zheng, H., Tandi, J., Zhou, J., Chee, M.W., 2017. Longitudinal brain structure and cognitive changes over 8 years in an East Asian cohort. NeuroImage 147, 852–860.

Rodrigue, K.M., Raz, N., 2004. Shrinkage of the entorhinal cortex over five years predicts memory performance in healthy adults. J. Neurosci. 24 (4), 956–963.

Lundervold, A.J., Vik, A., Lundervold, A., 2019. Lateral ventricle volume trajectories predict response inhibition in older age – a longitudinal brain imaging and machine learning approach. PLoS One 14 (4), e0207967.

Chandra, A., Dervenoulas, G., Politis, M., Initiative, A.D.N., et al., 2019. Magnetic resonance imaging in Alzheimer's disease and mild cognitive impairment. J. Neurol. 266 (6), 1293–1302.

Raz, N., 2000. Aging of the brain and its impact on cognitive performance: integration of structural and functional findings. In: Craik, F., Salthouse, T. (Eds.), The Handbook of Aging and Cognition. Lawrence Erlbaum Associates Publishers.

West, M.J., Coleman, P.D., Flood, D.G., Troncoso, J.C., 1994. Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. The Lancet 344 (8925), 769–772.

Thompson, P.M., Hayashi, K.M., De Zubicaray, G.I., Janke, A.L., Rose, S.E., Semple, J., Hong, M.S., Herman, D.H., Gravano, D., Doddrell, D.M., et al., 2004. Mapping hippocampal and ventricular change in Alzheimer disease. NeuroImage 22 (4), 1754–1766.

Scheltens, P., Leys, D., Barkhof, F., Huglo, D., Weinstein, H., Vermersch, P., Kuiper, M., Steinling, M., Wolters, E.C., Valk, J., 1992. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. J. Neurol. Neurosurg. Psychiatry 55 (10), 967–972.

Falahati, F., Westman, E., Simmons, A., 2014. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. J. Alzheimer's Dis. 41 (3), 685–708.

Shi, F., Liu, B., Zhou, Y., Yu, C., Jiang, T., 2009. Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: meta-analyses of MRI studies. Hippocampus 19 (11), 1055–1064.

Jack Jr., C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27 (4), 685–691.

Muller, S., Scealy, J.L., Welsh, A.H., et al., 2013. Model selection in linear mixed models. Stat. Sci. 28 (2), 135–167.

Ngufor, C., Van Houten, H., Caffo, B.S., Shah, N.D., McCoy, R.G., 2019. Mixed effect machine learning: a framework for predicting longitudinal change in hemoglobin A1c. J. Biomed. Inform. 89, 56–67.

Lei, B., Jiang, F., Chen, S., Ni, D., Wang, T., 2017. Longitudinal analysis for disease progression via simultaneous multi-relational temporal-fused learning. Front. Aging Neurosci. 9, 6.

Huang, L., Jin, Y., Gao, Y., Thung, K.-H., Shen, D., et al., 2016. Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. Neurobiol. Aging 46, 180–191.

Zhang, D., Shen, D., et al., 2012. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. PLoS One 7 (3), e33182.

Lim, B., van der Schaar, M., 2018. Forecasting Disease Trajectories in Alzheimer's Disease Using Deep Learning arXiv preprint arXiv:1807.03159.

Fischl, B., 2012. FreeSurfer. Neuroimage 62 (2), 774–781.

Trefler, A., Sadeghi, N., Thomas, A.G., Pierpaoli, C., Baker, C.I., Thomas, C., 2016. Impact of time-of-day on brain morphometric measures derived from T1-weighted magnetic resonance imaging. NeuroImage 133, 41–52.

Bernal-Rusiel, J.L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R., et al., 2013. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. NeuroImage 66, 249–260.

Mofrad, S.A., Lundervold, A.J., Vik, A., Lundervold, A.S., 2021. Cognitive and MRI trajectories for prediction of Alzheimer's disease. Sci. Rep. 11 (1), 1–10.

Gavidia-Bovadilla, G., Kanaan-Izquierdo, S., Mataró-Serrat, M., Perera-Lluna, A., et al., 2017. Early prediction of Alzheimer's disease using null longitudinal model-based classifiers. PLoS One 12 (1), e0168011.

Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N. T., Lenzo, N., Martins, R.N., Maruff, P., et al., 2009. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int. Psychogeriatr. 21 (4), 672–687.

Chepkoech, J.-L., Walhovd, K.B., Grydeland, H., Fjell, A.M., et al., 2016. Effects of change in FreeSurfer version on classification accuracy of patients with Alzheimer's disease and mild cognitive impairment. Hum. Brain Mapp. 37 (5), 1831–1841.

Gronenschild, E.H., Habets, P., Jacobs, H.I., Mengelers, R., Rozendaal, N., Van Os, J., Marcelis, M., 2012. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. PLoS One 7 (6), e38234.

Klauschen, F., Goldman, A., Barra, V., Meyer-Lindenberg, A., Lundervold, A., 2009. Evaluation of automated brain MR image segmentation and volumetry methods. Hum. Brain Mapp. 30 (4), 1310–1327.

Lindstrom, M.J., Bates, D.M., 1990. Nonlinear mixed effects models for repeated measures data. Biometrics 673–687.

Harrison, X.A., Donaldson, L., Correa-Cano, M.E., Evans, J., Fisher, D.N., Goodwin, C.E., Robinson, B.S., Hodgson, D.J., Inger, R., 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ 6, e4794.

West, B.T., Welch, K.B., Galecki, A.T., 2014. Linear Mixed Models: A Practical Guide Using Statistical Software. Chapman and Hall/CRC.

Bell, A., Jones, K., 2015. Age, period and cohort processes in longitudinal and life course analysis: a multilevel perspective. In: Burton-Jeangros, C., Cullati, S., Sacker, A., Blane, D. (Eds.), A Life Course Perspective on Health Trajectories and Transitions. Springer, Cham, pp. 197–213.

Seabold, S., Perktold, J., 2010. Statsmodels: econometric and statistical modeling with python. 9th Python in Science Conference.

Dietterich, T.G., 2000. Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems 1–15.

Saeys, Y., Abeel, T., Van de Peer, Y., 2008. Robust feature selection using ensemble feature selection techniques. Joint European Conference on Machine Learning and Knowledge Discovery in Databases 313–325.

Yue, L., Hu, D., Zhang, H., Wen, J., Wu, Y., Li, W., Sun, L., Li, X., Wang, J., Li, G., et al., 2021. Prediction of 7-year's conversion from subjective cognitive decline to mild cognitive impairment. Hum. Brain Mapp. 42 (1), 192–203.

Albert, M., Zhu, Y., Moghekar, A., Mori, S., Miller, M.I., Soldan, A., Pettigrew, C., Selnes, O., Li, S., Wang, M.-C., 2018. Predicting progression from normal cognition to mild cognitive impairment for individuals at 5 years. Brain 141 (3), 877–887.

Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S., Initiative, A. D.N., et al., 2013. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage: Clinical 2, 735–745.

Cui, R., Liu, M., Initiative, A.D.N., et al., 2019. RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. Comput. Med. Imaging Graph. 73, 1–10.

Cole, J.H., Franke, K., 2017. Predicting age using neuroimaging: innovative brain ageing biomarkers. Trends Neurosci. 40 (12), 681–690.

Walhovd, K.B., Fjell, A.M., Espeseth, T., 2014 Jun. Cognitive decline and brain pathology in aging-need for a dimensional, lifespan and systems vulnerability view. Scand. J. Psychol. 55, 244–254.

Nyberg, L., Pudas, S., 2019 01. Successful memory aging. Annu. Rev. Psychol. 70, 219–243.

Association, A.P., 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Pilgrim Press, Washington.

Di, X., Wolfer, M., Kühn, S., Zhang, Z., Biswal, B.B., 2019. Estimations of the weather effects on brain functions using functional MRI-a cautionary tale. bioRxiv 646695.

Hensel, A., Wolf, H., Dieterlen, T., Riedel-Heller, S., Arendt, T., Gertz, H.J., 2005. Morphometry of the amygdala in patients with questionable dementia and mild dementia. J. Neurol. Sci. 238 (1–2), 71–74.

Poulin, S.P., Dautoff, R., Morris, J.C., Barrett, L.F., Dickerson, B.C., et al., 2011. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Res. Neuroimaging 194 (1), 7–13.

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., Colliot, O., et al., 2020. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med. Image Anal. 101694.

Lundervold, A.S., Lundervold, A., 2019. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift fur medizinische Physik 29, 102–127.

Sperling, R., 2011. The potential of functional MRI as a biomarker in early Alzheimer's disease. Neurobiol. Aging 32, S37–S43.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N. C., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131 (3), 681–689.

Lajoie, I., Nugent, S., Debacker, C., Dyson, K., Tancredi, F.B., Badhwar, A., Belleville, S., Deschaintre, Y., Bellec, P., Doyon, J., et al., 2017. Application of calibrated fMRI in Alzheimer's disease. NeuroImage: Clinical 15, 348–358.

Doan, N.T., Engvig, A., Persson, K., Alnæs, D., Kaufmann, T., Rokicki, J., Córdova-Palomera, A., Moberget, T., Brækhus, A., Barca, M.L., et al., 2017. Dissociable diffusion MRI patterns of white matter microstructure and connectivity in Alzheimer's disease spectrum. Sci. Rep. 7, 45131.

Kim, J., Basak, J.M., Holtzman, D.M., 2009. The role of apolipoprotein E in Alzheimer's disease. Neuron 63 (3), 287–303.

Safieh, M., Korczyn, A.D., Michaelson, D.M., 2019. ApoE4: an emerging therapeutic target for Alzheimer's disease. BMC Med. 17 (1), 1–17.

Janelidze, S., Stomrud, E., Smith, R., Palmqvist, S., Mattsson, N., Airey, D.C., Proctor, N. K., Chai, X., Shcherbinin, S., Sims, J.R., et al., 2020. Cerebrospinal fluid p-tau217 performs better than p-tau181 as a biomarker of Alzheimer's disease. Nat. Commun. 11 (1), 1–12.

Holleran, L., Kelly, S., Alloza, C., Agartz, I., Andreassen, O.A., Arango, C., Banaj, N., Calhoun, V., Cannon, D., Carr, V., et al., 2020. The relationship between white matter microstructure and general cognitive ability in patients with schizophrenia and healthy participants in the ENIGMA consortium. Am. J. Psychiatry pp. appi-ajp.

PAPER B

COGNITIVE AND MRI TRAJECTORIES FOR PREDICTION OF ALZHEIMER'S DISEASE

Mofrad, Samaneh A., Lundervold, Astri J., Vik, Alexandra and Lundervold, Alexander S. 22 January, 2021. Cognitive and MRI trajectories for prediction of Alzheimer's disease. *Scientific Reports*. Nature Publishing Group.

scientific reports

OPEN

Cognitive and MRI trajectories for prediction of Alzheimer's disease

Samaneh A. Mofrad[1,3]✉, Astri J. Lundervold[2], Alexandra Vik[3] & Alexander S. Lundervold[1,3]

The concept of Mild Cognitive Impairment (MCI) is used to describe the early stages of Alzheimer's disease (AD), and identification and treatment before further decline is an important clinical task. We selected longitudinal data from the ADNI database to investigate how well normal function (HC, n= 134) vs. conversion to MCI (cMCI, n= 134) and stable MCI (sMCI, n=333) vs. conversion to AD (cAD, n= 333) could be predicted from cognitive tests, and whether the predictions improve by adding information from magnetic resonance imaging (MRI) examinations. Features representing trajectories of change in the selected cognitive and MRI measures were derived from mixed effects models and used to train ensemble machine learning models to classify the pairs of subgroups based on a subset of the data set. Evaluation in an independent test set showed that the predictions for HC vs. cMCI improved substantially when MRI features were added, with an increase in F_1-score from 60 to 77%. The F_1-scores for sMCI vs. cAD were 77% without and 78% with inclusion of MRI features. The results are in-line with findings showing that cognitive changes tend to manifest themselves several years after the Alzheimer's disease is well-established in the brain.

Ageing is associated with cognitive changes characterised by phenotypic diversity in both pace and severity. This diversity is a result of the many biological and life-style factors influencing an individual throughout his or her life-time[1,2]. Some individuals preserve their cognitive function into old age, so-called "superagers"[3], while others experience a decline at a younger age due to a neurodegenerative disease[4]. Along this wide dimension of cognitive function, it becomes difficult to define the fine line between normal and pathological ageing.

Alzheimer's disease (AD) is a common neurodegenerative disease characterised by a cognitive impairment that gradually worsens over time[5]. A lot of effort has been put into the identification and development of treatment options that can stop this degenerative process at an early stage. Early on, the cognitive symptoms tend to be minor and the condition is referred to as a Mild Cognitive Impairment (MCI)[6]. Not all patients with MCI will develop AD. Although studies have shown that a patient with MCI has up to a tenfold increased risk to develop the disease[4,7], a subgroup of individuals with MCI are left with a stable condition or may even revert to normal function[8]. The search for predictors of conversion from MCI to AD is therefore an important field of research[6,9].

Impaired performance on psychometric tests of memory function[10,11] and on more global measures of cognitive function[9] have been recognized as early cognitive predictors of AD. However, this impairment tend not to be uncovered until years after the condition is well-established in the brain[12]. This is documented by several previous studies relating early changes in cognitive function to changes in specific regions and structures of the brain, including an expansion of the ventricles and volume loss in the hippocampus and entorhinal cortex[13,14]. A more precise prediction of AD is therefore expected if information from results on cognitive tests are combined with information from magnetic resonance imaging (MRI) of the brain[15,16].

The present study was motivated by the challenge to predict AD at an early stage of the disease. Based on data available from the Alzheimer's Disease Neuroimaging Initiative (ADNI) we investigated how well a set of machine learning models could predict conversion from normal function through MCI to AD. In a first set of analyses we defined features characterising longitudinal changes in memory function (Rey Auditory Learning Test (RAVLT))[11] and in a more global measure of cognitive function (ADAS-Cog-13 (ADAS13))[9,17]. Expecting more precise predictions by including information from MRI examinations[15,16], we investigated the add-on effect of including morphometric brain measures associated with memory function (entorhinal cortex and hippocampus[14]) and a global measure of cognitive function (the volume of the ventricles as a proxy for a global

[1]Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Pb. 7030, Bergen 5020, Norway. [2]Department of Biological and Medical Psychology, University of Bergen, Bergen, Norway. [3]MMIV, Department of Radiology, Haukeland University Hospital, Bergen, Norway. ✉email: sam@hvl.no

Labels in ADNI and our longitudinal labels		
Labels	**Subgroup**	**Description**
ADNI	CN	Cognitively normal at visit
	MCI	Mild cognitive impairment at visit
	Dementia	Alzheimer's disease at visit
Our study	HC	CN at all visits
	cMCI	Initially CN but later converted to MCI
	sMCI	MCI at all visits
	cAD	Initially MCI but later converted to AD
	sAD[a]	Dementia at all visits

Table 1. The original ADNI labels and the longitudinal labels used in the present study. [a] The sAD subgroup was not included in the present study as we focused on converters.

Information about the subgroups					
Subgroups	**#Subjects**	**#Images**	**Gender (f/m)**	**Average #visits (mri/cog)**	**Average time (mri/cog)**
HC	134	642	57/77	4.8/6.0	0.58/0.78
cMCI	134	731	55/79	5.5/7.0	0.62/0.80
sMCI	333	1696	143/190	5.1/6.0	0.48/0.63
cAD	333	1871	130/203	5.6/6.5	0.53/0.62
ALL	934	4904	385/549	5.3/6.3	0.54/0.67

Table 2. Total number of subjects, T1-weighted MR images, and gender distribution within each of the four subgroups. The table also shows the average number of MRI scans (mri) and cognitive tests (cog) per subject, available in each subgroup, and the average time (in years) between the MRI scans and cognitive tests per subgroup.

tissue loss[18]). More specifically, we used a pipeline proposed by Mofrad et al.[19] based on a combination of mixed effects and machine learning models for analysis of longitudinal data. This approach is useful when faced with a set of subjects with a varying number of scans and test results, examined at different time intervals. This is a common challenge in longitudinal studies, including studies based on the ADNI dataset.

Materials and methods

Data set.

Data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD, with an overall goal to validate biomarkers for use in clinical treatment trials for patients with AD. The study was approved by the Institutional Review Boards at each ADNI site (see full list here: http://adni.loni.usc.edu). Informed consent was obtained from all subjects prior to enrollment. All methods were carried out in accordance with relevant guidelines and regulation. The present study was approved by ADNI Publication Committee (ADNI DPC).

In the present study we included subjects from the ADNI dataset defined as cognitively normal (CN) or as patients with an MCI or AD diagnosis. According to the ADNI protocol, MCI was defined if a participant or caregivers reported cognitive problems, if the patient showed impairment on the logical memory-II subtest from Wechsler memory scale-R, a mini-mental state examination score equal to or above 24, and a clinical dementia rating = 0.5. None of the participants with MCI should meet the diagnostic criteria for dementia. AD was diagnosed according to the NINCDS-ADRDA Alzheimer's Criteria for probable AD (see http://adni.loni.usc.edu/methods/documents for details).

We defined four subgroups from the ADNI sample, with a restriction to subjects with MRI scans at least at two time-points and results on two selected psychometric tests of cognitive function. We labelled subjects as *healthy controls* (HC) if they were classified as CN at all ADNI visits. The subjects who converted from CN to MCI during the observation period were labelled *converted MCI* (cMCI). Subjects who were defined with MCI at all visits were labelled *stable MCI* (sMCI) and those converting from MCI to AD were labelled *converted AD* (cAD) (see Table 1). We balanced the number of subjects in each pair of subgroups, (HC, cMCI) and (sMCI, cAD), controlling for age and gender, and ended up with a total of 934 subjects. See Tables 2 and 3 for details.

Cognitive and MRI measures.

The *RAVLT* was included as a measure of memory function. In this test, the participants are asked to recall words from a list of 15 nouns immediately after each of five learning trials and after a short and a long delay. Two measures known to be sensitive to cognitive changes in patients with AD[11] were included in the present study: *Immediate recall* (RAVLT-Im): the number of correct responses across the immediate recall of the five

	Subgroups				p-values
Variables	**HC**	**cMCI**	**sMCI**	**cAD**	**(HC-cMCI)/(sMCI-cAD)**
Age					
Train (f/m)	77.4±7/77.7±7	75.2±7/77±7	75.2±8/77±7	75.6±8/77.6±7	(∗ ∗ ∗/−)/(− /−)
Test (f/m)	77.2±7/78.3±6	76.6±9/77.2±6	72.6±6/76±7	72.3±8/77.2±7	(−/−)/(−/−)
Education					
Train (f/m)	15.1±3/ 17.5±2	16±2/17±2	15.6±3/16.5±3	15.1±3/16.2±3	(∗ ∗ ∗/∗ ∗ ∗)/(∗∗/−)
Test (f/m)	16.1±3/17.2±3	17±2/15.8±4	13.1±3/15.8±3	15.9±3/16.4±3	(∗/∗∗)/(∗ ∗ ∗/∗)
RAVLT-Im					
Train (f/m)	47.4±10/43.8±11	47.3±10/39.6±10	38.7±11/33.2±10	29.4±9/28.2±7	(−/∗)/(∗/∗)
Test (f/m)	48.3±8/40.8±8	51.3±14/35.5±7	38.4±12/32.8±10	30.1±10/26.4±6	(−/∗ ∗ ∗)/(∗ ∗ ∗/∗ ∗ ∗)
RAVLT-PF					
Train (f/m)	30.8±27/36±30	33.1±26/43±27	54.7±33/58.4±32	81.8±28/77.1±27	(−/∗∗)/(∗ ∗ ∗/∗ ∗ ∗)
Test (f/m)	31.4±25/35.9±25	33.8±28/48.3±29	51.9±35/56.1±32	81.6±30/81.4±25	(−/∗∗)/(∗ ∗ ∗/∗ ∗ ∗)
ADAS13					
Train (f/m)	8.2±4/9.9±5	8.6±4/10.9±5	14.3±7/15.3±7	21.8±7/19.7±6	(−/∗∗)/(∗ ∗ ∗/∗ ∗ ∗)
Test (f/m)	8.1±4/8.1±3	8.3±5/12.2±3	13.8±8/15.7±6	22.1±7/20.4±6	(−/∗ ∗ ∗)/(∗ ∗ ∗/∗ ∗ ∗)

Table 3. Means and standard deviations of age, education level, and scores on the included cognitive tests for each subgroup, given separately for the training and test sets. The information for the converted subgroups (cMCI and cAD) is calculated after removing the measurements from point of conversion and onward. The p-values for pairs of subgroups are presented separately for females and males; ∗: p < .05; ∗∗: p < .01; ∗ ∗ ∗: p < .001; −non-significant at 0.05 level.

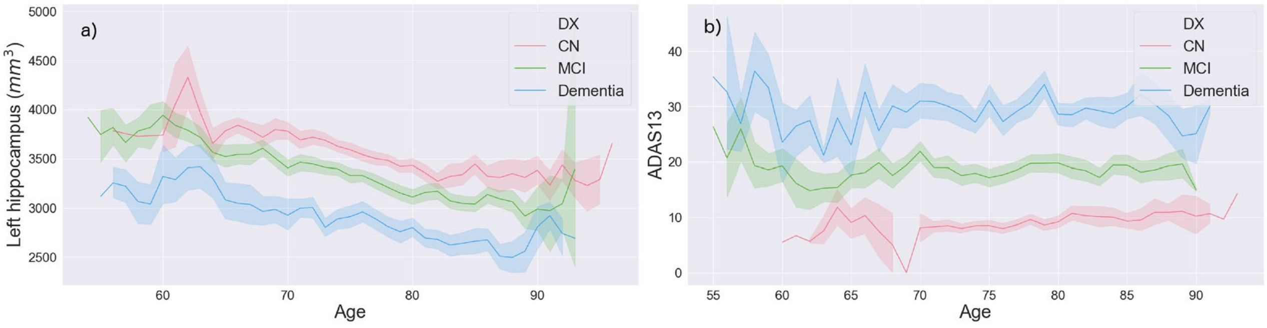


Figure 1. Mean values for (**a**) the volume of the left hippocampus, and (**b**) the ADAS13 score over age, based on the cross-sectional ADNI labels.

learning trials; *percent forgetting* (RAVLT-PF): the score on the fifth learning trial minus the score on the long delayed recall, divided by the score obtained on the fifth learning trial. The lower the scores, the more severe impairment of cognitive function.

The *ADAS13* was included as a global measure of cognitive function. ADAS13 is a test battery developed to assess severity of cognitive impairment associated with AD and includes subtests and clinical evaluations assessing memory function, reasoning, language function, orientation and praxis. The ADAS13 is a modified version of the original ADAS-Cog-11[20], adding a cancellation task and a delayed free recall task[21]. The higher the scores, the more severe impairment of cognitive function.

We used Freesurfer v.6.0[22] to derive measures from the T1-weighted MR images, extracting the lateral ventricle volumes, the volumes of the hippocampus and the thickness of the entorhinal cortex in the left and right hemisphere. To reduce the effect of individual and gender differences in brain sizes, the volumes were normalized using a total intracranial volume measure estimated by Freesurfer (eTIV).

Figure 1 shows the age-dependent volume changes in the hippocampus (left hemisphere) and ADAS13 test scores across age. The severity of the volume loss and impairment on the ADAS13 are gradually increased from the HC through MCI to AD in the ADNI dataset. Figure 2 illustrates that the more severe scores in patients with AD compared to the other groups are found in both males and females, with a trend towards higher scores (i.e., better results) in females than males on the memory test in the CN and the MCI groups. Means and standard deviations for the RAVLT and the ADAS13 test scores are presented in Table 3.

Features. To construct subject specific trajectories for each measure we used linear mixed effects models[23,24], a class of models able to produce regression models from dependent variables[25]. Our models are based on the one presented in[24] and similar to the ones employed in our previous works[19,26]. As the ventricles show quadratic

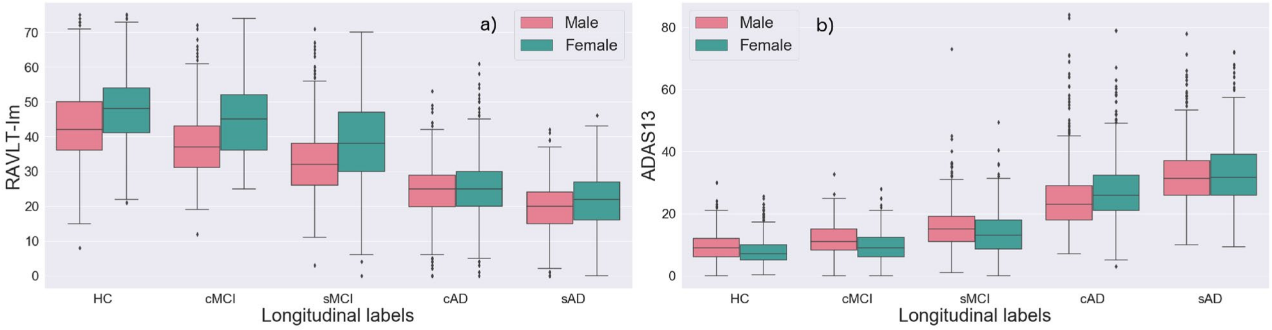


Figure 2. Box plot showing the gender specific results on RAVLT immediate recall and the ADAS13 for each of the longitudinal labels defined for the present study (Table 1).

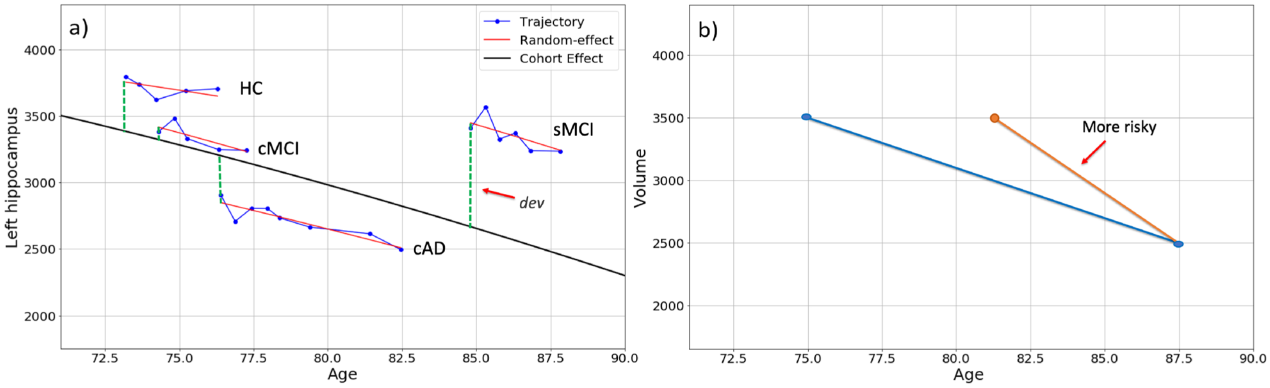


Figure 3. (**a**) Trajectories of age-related changes in a volumetric MRI measure (i.e., left-hippocampus) and random effects in four subjects for each of the four subgroups. The distance between the cohort effect and random effect (*dev*) of each subject (the green vertical lines) was included as one of the features in our statistical models. (**b**) The time-span was different between the participants in the present study. The change in ROI volume may therefore be the same for a participant with a short and long participation time, here illustrated by the red and blue line, respectively. The *d-slope* feature is included to capture this phenomenon[19].

cohort behaviour (Fig. 5), likely caused by the accumulation of cerebrospinal fluid due to atrophy in multiple brain regions, we used linear mixed effects models both with and without a quadratic covariate term:

$$\mathrm{M}_{\mathrm{ij}}^{\mathrm{c}} = \underbrace{\beta_0^c + \beta_1^c \mathrm{Age}_{\mathrm{ij}}}_{\text{fixed effect}} + \underbrace{b_{0i}^c + b_{1i}^c \mathrm{Age}_{\mathrm{ij}} + \epsilon_{\mathrm{ij}}^{\mathrm{c}}}_{\text{random effect}}, \tag{1}$$

$$\mathrm{M}_{\mathrm{ij}}^{\mathrm{c}} = \underbrace{\beta_0^c + \beta_1^c \mathrm{Age}_{\mathrm{ij}} + \beta_2^c \mathrm{Age}_{\mathrm{ij}}^2}_{\text{fixed effect}} + \underbrace{b_{0i}^c + b_{1i}^c \mathrm{Age}_{\mathrm{ij}} + \mathrm{b}_{2i}^c \mathrm{Age}_{\mathrm{ij}}^2 + \epsilon_{\mathrm{ij}}^{\mathrm{c}}}_{\text{random effect}}, \tag{2}$$

where c denotes the brain region or cognitive test score, M_{ij}^c is the measurement of volume of region c or score of cognitive test c for subject $i = 1, \ldots, N$ at referral $j = 1, \ldots, n_i$. n_i is the number of MRI scans or cognitive tests for subject i. Age_{ij} is age of subject i at referral j. Age is the only predictor variable in the mixed model. The β_0^c, β_1^c, and β_2^c are fixed effect parameters while b_{0i}^c, b_{1i}^c, and b_{2i}^c are random effect parameters. ϵ_{ij}^c denotes the random residual errors.

For constructing the mixed effects models we used the `mixedlm` function from the `statsmodels` Python library (v. 0.9.0). For each cognitive and MRI measure we derived the following features for each subject: (i) *r-slope*: the model-based random effects slope, thus taking the cohort effects for all subjects, and duration of study for each individual into account (the slope of the red lines in Fig. 3a). For both the linear model (Eq. 1) and the quadratic mixed models (Eq. 2), *r-slope* is b_{1i}^c, but for the Eq. 2 we used the coefficient of the quadratic term, b_{2i}^c, as an additional feature. (ii) *dev*: the distance (deviance) between the random effect line and the fixed effect line at the first time point ($\mathrm{M}_{\mathrm{i1}} - (\beta_0 + \beta_1 \mathrm{Age}_{\mathrm{i1}})$), thus taking the results at entry point into account (green dashed lines in Fig. 3a); (iii) *d-slope*: the slope obtained by dividing the difference of the measure at the first and last measurements by the duration between them, i.e. the slope of change from the first to the last measurement:

$$\mathrm{d-slope_i} = \frac{\mathrm{M_{in_i}} - \mathrm{M_{i0}}}{\mathrm{Age_{in_i}} - \mathrm{Age_{i0}}} \tag{3}$$

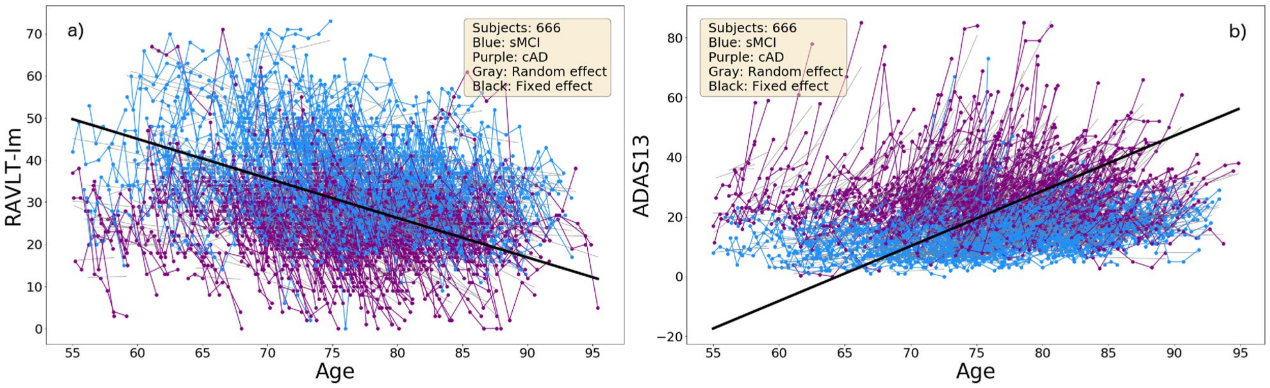


Figure 4. The trajectories for performances on the RAVLT-Im test (**a**) and the ADAS13 test (**b**), with age at testing on the x-axis. The thick black curve is the cohort regression line, and thin grey lines are random effects for each subject. Severity of impairment is reflected by a lower score on the RAVLT test and a higher score on the ADAS13.

where M_{i0} and M_{in_i} are the measurement at the first and last visits for subject *i*, respectively. This slope was used because identical changes in brain measurements or test scores can occur over different time spans, and the period of participation in the study varies for different subjects[19] (Fig. 3b). We also added *sex* and age at last visit (*Current-Age*) before conversion, if applicable (MCI in cMCI, and AD in cAD), as features for the predictive models.

Machine learning models and feature importance. We investigated the following experiments:

1. Classifying subjects with stable MCI (sMCI, n = 333, f/m = 143/190) vs. those who converted from MCI to AD (cAD, n = 333 , f/m = 130/203).
2. Classifying healthy controls (HC, n = 134 , f/m = 57/77) vs. those who converted from being a healthy control to MCI (cMCI, n = 134 , f/m = 55/79).

No features based on information from the point of conversion and onward were made available to the models, as they were tasked with making predictions about future diagnostic status.

In mixed effects models each group (i.e. each subject) influence the fixed effect model, and therefore impacts all the other subjects' trajectories[27]. To avoid data leakage caused by the resulting influence on the derived features, we put aside a test set containing 20% of the subjects before creating the mixed effects models. We balanced the number of subjects in each class and controlled for gender and age. No subjects were present in both the train- and test set.

We trained an ensemble model based on a soft voting strategy, i.e. based on a weighted vote taking the models assigned probabilities into account, containing the following five models: logistic regression, support vector machine, K nearest neighbors, random forest, and a gradient boosting model. We used an ensemble approach as this tend to result in more robust classifiers that are less reliant on specific properties in the data set when compared to single classifiers[28,29]. We used confusion matrices, precision, recall and F_1 scores to assess our models during development and hyperparameter selection, using subject-level, leave-one-out cross-validation on the training set. For each model we set up a grid search through hyperparameters to select models that generalized well. For the logistic regression model we evaluated whether to include *l*2 penalty and the strength of regularization. For the support vector machine model we assessed various kernels (polynomial, sigmoid and radial basis function), the kernel coefficient and regularization parameter. For the K nearest neighbor model we tried multiple combinations of the number of neighbors and distance metrics. For the random forest model we searched for a good combination of the number of trees and the maximum tree depth allowed, while for the gradient boosting model we searched through both complexity and sampling parameters. To ensure fair comparison among the models trained on different sets of features, we ran new grid searches for each feature set.

To evaluate the feature importance in the classification model, we used permutation importance, also called mean decrease accuracy, as implemented in the `ELI5` Python library. This is a data-driven approach to feature importance, based on measuring the decrease in model accuracy when randomly shuffling each feature separately multiple times (we used five trials for each feature). The idea is that the negative impact on performance of permuting an important feature is larger than for less important features[30].

Results

Experiment 1: Prediction of sMCI vs. cAD. The change in performances on the RAVLT-Im and ADAS13 tests are illustrated in Fig. 4. Note the age-related decline in both the sMCI and the cAD subgroups, with the most severe impairments shown within the cAD group.

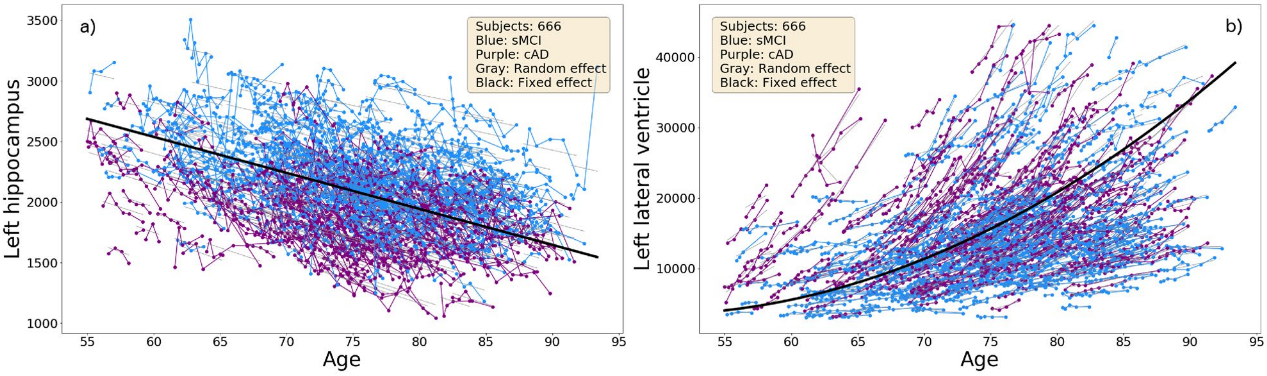


Figure 5. The trajectories for the normalized volumes of the hippocampus and the lateral ventricle in the left hemisphere with age at scan at the x-axis. The thick black curve is the cohort regression line, and the thin grey lines are random effects for each subject.

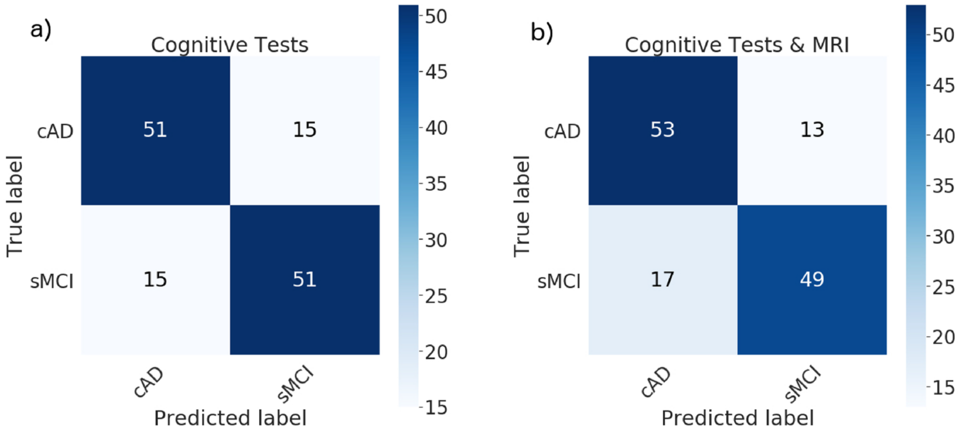


Figure 6. Confusion matrices for classification of sMCI vs. cAD from the cognitive features (**a**) and the combination of MRI and cognitive features (**b**).

Figure 5 illustrates age-related tissue loss in the brain, with an almost linear shrinkage of the hippocampus volumes (Fig. 5a) and a non-linear increase in the volume of the lateral ventricle (Fig. 5b). Overall, the most extensive losses are found among subjects in the cAD subgroup.

Inclusion of the cognitive trajectory features (*r-slope*, *dev* and *d-slope* for each test measure) in the ensemble model gave 77% for the accuracy, precision, recall and the F_1 scores. These scores changed to 77%, 76%, 80% and 78%, respectively, when the longitudinal MRI features were added. The confusion matrices in Fig. 6 show a misclassification rate of 23% for the subjects in both the cAD and the sMCI group when only the cognitive features were included, with a reduction to 20% for the cAD subgroup and an increase to 26% in the sMCI subgroup when the MRI features were added.

To further study these findings we performed a 15-fold cross validation experiment on the training data set, controlling for labels, age, and gender in the hold-out folds. The classifier trained on only cognitive features obtained a mean accuracy of 76% ± 4% and the MRI features resulted in mean accuracy of 77% ± 3.7%. Note that the models tested on the original hold-out test set were optimized based on leave-one-out cross validation over the entire training data set.

The part a) of Fig. 7 shows the weights of the features in our model classifying sMCI vs. cAD. The model-based random slope (*r-slope*) of the ADAS13 trajectory provided the strongest weight among the cognitive features. When the MRI features were included in the analysis, the weight of ADAS13 decreased substantially, and became stronger for features characterising the entorhinal cortex (*d-slope* and *dev-RH entorhinal*).

Experiment 2: Prediction of HC vs. cMCI. With the longitudinal cognitive features as inputs to our ensemble model, we obtained an accuracy, precision, recall and F_1 score of 62%, 62%, 58% and 60%, respectively. Adding the MRI features increased the accuracy, precision, recall and F_1 scores to 77% for all. The part a) of the confusion matrix in Fig. 8 shows a somewhat lower misclassification rate for HC subjects (35%) than for cMCI (42%) subjects when only the cognitive features were included in the analysis. The rate decreased to 23% for both subgroups when the MRI features were added (Fig. 8b).

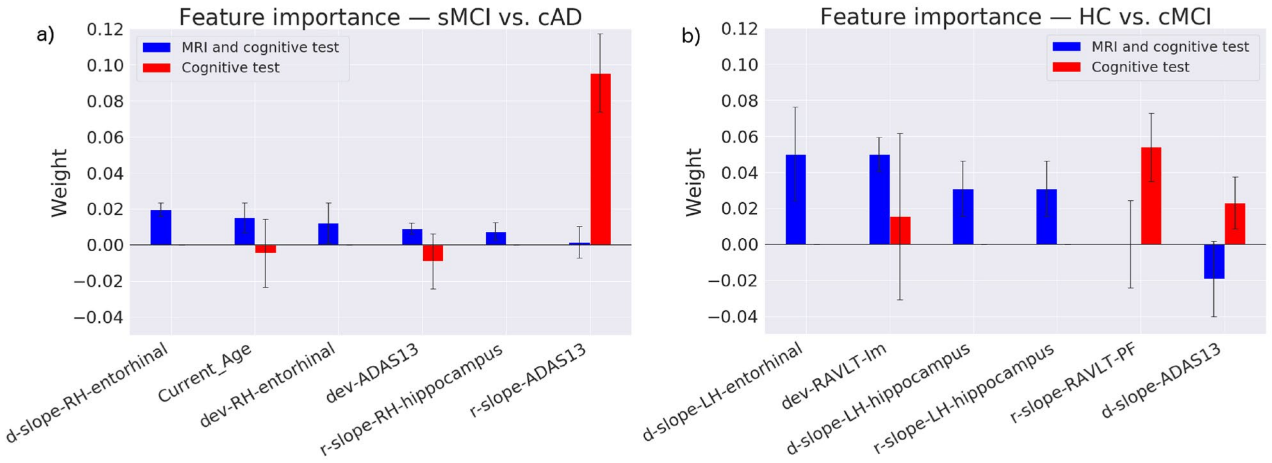


Figure 7. Feature weights when classifying sMCI vs. cAD (**a**) and HC vs. cMCI (**b**), based on cognitive features (in red) and the combination of MRI and cognitive features (in blue). For convenience, the plots only show a selection of the most important features after adding the MRI features to the analyses. Weights near zero and features for which the permutation importance had standard deviations greater than the estimated mean weight are not plotted. The most important features, when predicting from only the cognitive tests, were kept in the plot to illustrate the main changes observed after adding the MRI features.

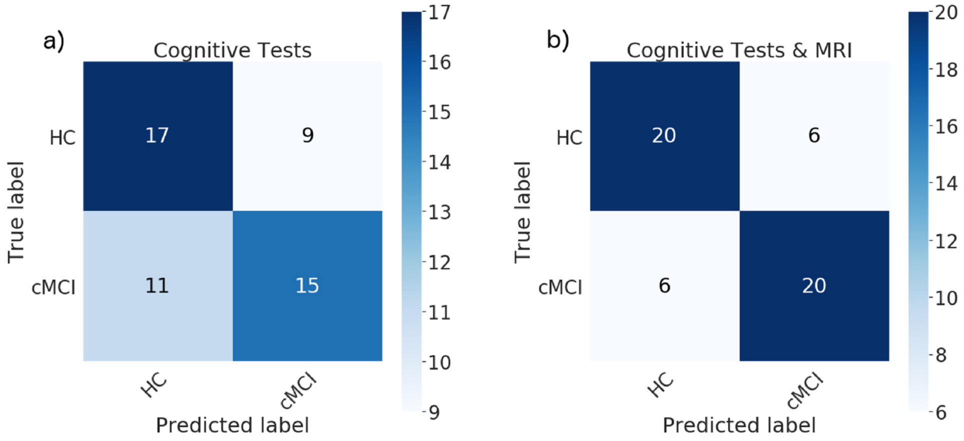


Figure 8. Confusion matrices for classifying HC vs. cMCI from cognitive features (**a**) and the combination of cognitive and MRI features (**b**).

To assess the robustness of the findings we again performed a 15-fold cross validation experiment on the training data. The classifier trained on only cognitive features gave a mean accuracy of 56% ± 6%, while the MRI features resulted in mean accuracy of 61% ± 5.7%.

The part (b) of Fig. 7 shows the feature importance for our model classifying HC vs. cMCI. The model-based random slope (*r-slope*) from a measure of memory function (RAVLT-PF) provided the strongest weight among the cognitive features. When the MRI measures were included, the *d-slope* of the entorhinal cortex in the left hemisphere and a measures of immediate memory function (*dev-RAVLT-im*) showed the strongest weights.

Discussion

The present study used mixed effects models to define features characterising individual trajectories of change in a set of cognitive and MRI measures. These features were then used as predictors to classify subgroups with stable MCI (sMCI) vs. converters to AD (cAD) in one experiment, and to classify subgroups of healthy controls (HC) vs. converters to MCI (cMCI) in a second experiment. Visual inspections showed an age-related decline in cognitive performance and volumetric MRI measures in all subgroups. Using the features to train ensemble machine learning models gave classifications that were clearly better than chance level. For the prediction of sMCI vs. cAD, the mean classification F_1-score was 77% when only the features characterising the trajectories of cognitive changes were included, with only one percentage point improvement when the MRI features were added. When restricted to the cognitive features, the model-based slope of the ADAS13 trajectory was given a relatively strong weight, while it was dramatically reduced and outperformed by features characterising the volume change in the entorhinal cortex when information from MRI was added. For the HC vs. cMCI predictions, the F_1-score was substantially improved from 60% to 77% when the MRI features were included. Among the cognitive features, a feature characterising change in memory function was given the strongest weight, followed by ADAS13. When the MRI features were added, information about the changes in the volume of the entorhinal cortex, hippocampus and the immediate memory function were given the strongest weights. The confusion matrices showed results above chance level, with the largest drop in misclassification rate when both the cognitive and MRI features were included.

The results confirmed the expected age-related change in cognitive function. Furthermore, the weight given to longitudinal features of memory function (in the HC vs. cMCI experiment) supports the sensitivity of memory tests to the early symptoms of a path leading towards a neurodegenerative disorder[10,11], and that symptoms of an amnesic MCI may indicate a high risk of a path towards AD[6]. In a stage closer to an AD diagnosis, the results on a more global measure of cognitive function (ADAS13)[9] are given stronger weight. Still, the contribution from MRI measures was substantial when classifying HC vs. cMCI. The design of the present study was inappropriate for identifying the exact time-point where information about MRI measures would increase the accuracy of the prediction. However, the results are still in line with studies showing that cognitive changes associated with AD tend to manifest themselves several years after the condition is well established in the brain[12]. The importance of the trajectory of change in the volume of the entorhinal cortex is also worth a comment. Entorhinal cortex acts like a relay station, with widespread connections to cortical and subcortical areas[31]. Several studies have documented that volume changes in the entorhinal cortex can be detected in an early stage of AD, and that there are strong correlations between different parts of the entorhinal cortex and memory function[32]. The present study should therefore be followed by studies on the predictive values of subcomponents of entorhinal, hippocampus and other related brain structures.

Although we obtained correct classifications above chance level, the misclassifications are too high to enable prediction on an individual level from the selected features. For converters to MCI, consideration should be given to the high number of individuals misclassified as healthy controls when the algorithms were based only on cognitive features. This illustrates the challenge in defining the fine line between healthy and pathological cognitive ageing, and the phenotypic diversity characterising the group of patients with MCI[1,2,33]. Furthermore, it may also reflect a limitation of the ADNI protocol. Although MCI is defined from the presence of subjective memory complaints, objective memory impairment, normal general cognitive function and intact activities of daily living/absence of dementia, studies have described heterogeneous subtypes, including a subgroup demonstrating intact cognitive function[34] and MRI findings[35]. The prediction was more accurate for classification of patients converting to AD than in those with a stable MCI. This indicates the challenge in classifying an individual as AD, a diagnosis that is only definite after a post-mortem confirmation[5]. Future studies including such a definite outcome measure are therefore warranted.

The high number of participants included in the present study and the inclusion of predictive models and methods from modern machine learning frameworks[36] are main strengths of the present study. The results in the study must, however, be interpreted in the light of several limitations. As already mentioned, this includes how we defined the subgroups. Inclusion of a small number of cognitive and MRI measures among the ones available in the ADNI dataset is another limitation. We have not provided sufficient information to specify whether the impairments in the MCI group affect single or multiple cognitive domains. And even the ADNI dataset miss out some important biomarkers[37] and information about cognitive reserve factors (e.g.[38,39]), factors that certainly are essential to understand the phenotypic diversity of trajectories from normal function to AD. The results are also restricted by our analytic approach. The choice of models not only influence the predictive performance, but also the feature weights indicating feature importance. Furthermore, as the method used to assess feature importance is based on permuting single features, it doesn't give a precise way to assess how combinations of features are weighed by the models. Finally, information about mean time between MRI scans and cognitive testing and number of visits, presented in Table 2, was not controlled for in the statistical models.

Conclusion

We showed that a set of mixed effects-derived features from psychometric tests of cognitive function and an MRI examination gave predictions of healthy controls vs. MCI and stable MCI vs. AD that were above chance level. The results confirmed the importance of early changes in memory function and the role of entorhinal cortex as an imaging-based biomarker of normal and pathological ageing in older adults. Our major contributions are the application of (i) measures from the rich ADNI dataset, (ii) features defining trajectories of change in relevant cognitive and MRI measures, and (iii) a data-driven machine learning approach to assess the measures' weights in classification models. Future studies should further investigate this avenue of brain-behaviour relationships in older age. They should consider inclusion of the wider range of genetic[40] and environmental[41] variables, and thus probably reduce the misclassifications shown in the present study, as well as other predictive models and methods within modern machine learning frameworks[36,42].

Received: 17 June 2020; Accepted: 17 November 2020
Published online: 22 January 2021

References

1. Walhovd, K. B., Fjell, A. M. & Espeseth, T. Cognitive decline and brain pathology in aging-need for a dimensional, lifespan and systems vulnerability view. *Scand. J. Psychol.* **55**, 244–54 (2014).
2. Nyberg, L. & Pudas, S. Successful memory aging. *Annu. Rev. Psychol.* **70**, 219–243 (2019).
3. Rogalski, E. J. *et al.* Youthful memory capacity in old brains: Anatomic and genetic clues from the Northwestern SuperAging Project. *J. Cogn. Neurosci.* **25**, 29–36 (2013).
4. Petersen, R. C. Mild cognitive impairment or questionable dementia?. *Arch. Neurol.* **57**, 643–644 (2000).

5. Association, A. P. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* (Pilgrim Press, Washington, 2013).
6. Petersen, R. C. Mild cognitive impairment as a diagnostic entity. *J. Intern. Med.* **256**, 183–194. https://doi.org/10.1111/j.1365-2796.2004.01388.x (2004).
7. Boyle, P. A., Wilson, R. S., Aggarwal, N. T., Tang, Y. & Bennett, D. A. Mild cognitive impairment: Risk of Alzheimer disease and rate of cognitive decline. *Neurology* **67**, 441–445 (2006).
8. Bennett, D. A. *et al.* Natural history of mild cognitive impairment in older persons. *Neurology* **59**, 198–205 (2002).
9. Yagi, T. *et al.* Identification of prognostic factors to predict cognitive decline of patients with early Alzheimers disease in the Japanese Alzheimers Disease Neuroimaging Initiative study. *Alzheimers Dement.* **5**, 364–373 (2019).
10. Belleville, S. *et al.* Neuropsychological measures that predict progression from mild cognitive impairment to Alzheimers type dementia in older adults: A systematic review and meta-analysis. *Neuropsychol. Rev.* **27**, 328–353 (2017).
11. Moradi, E., Hallikainen, I., Hänninen, T., Tohka, J. & Initiative, A. D. N. Reys auditory verbal learning test scores can be predicted from whole brain MRI in Alzheimers disease. *NeuroImage. Clin.* **13**, 415–427 (2017).
12. Jack, C. R. Jr. & Holtzman, D. M. Biomarker modeling of Alzheimers disease. *Neuron* **80**, 1347–58 (2013).
13. Leong, R. L. *et al.* Longitudinal brain structure and cognitive changes over 8 years in an East Asian cohort. *Neuroimage* **147**, 852–860 (2017).
14. Raz, N. Decline and compensation in aging brain and cognition: Promises and constraints preface. *Neuropsychol. Rev.* **19**, 411–414 (2009).
15. Zandifar, A. *et al.* MRI and cognitive scores complement each other to accurately predict Alzheimer dementia 2 to 7 years before clinical onset. *NeuroImage. Clin.* **25**, 102121 (2020).
16. Moreland, J. *et al.* Validation of prognostic biomarker scores for predicting progression of dementia in patients with amnestic mild cognitive impairment. *Nucl. Med. Commun.* **39**, 297–303 (2018).
17. Mohs, R. C. *et al.* Development of cognitive instruments for use in clinical trials of antidementia drugs: Additions to the alzheimers disease assessment scale that broaden its scope. *Alzheimer Dis. Assoc. Disord.* **11**, 13–21 (1997).
18. Lundervold, A. J., Vik, A. & Lundervold, A. Lateral ventricle volume trajectories predict response inhibition in older age-A longitudinal brain imaging and machine learning approach. *PLoS One* **14**, e0207967 (2019).
19. Mofrad, S. A., Lundervold, A. & Lundervold, A. S. A predictive framework based on brain volume trajectories enabling early detection of Alzheimer's disease (2020). Submitted.
20. Kueper, J. K., Speechley, M. & Montero-Odasso, M. The Alzheimers disease assessment scale-cognitive subscale (ADAS-Cog): Modifications and responsiveness in pre-dementia populations A narrative review. *J. Alzheimers Dis. JAD* **63**, 423–444 (2018).
21. Skinner, J. *et al.* The Alzheimers disease assessment scale-cognitive-plus (ADAS-Cog-Plus): An expansion of the ADAS-Cog to improve responsiveness in MCI. *Brain Imaging Behav.* **6**, 489–501 (2012).
22. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
23. Müller, S. *et al.* Model selection in linear mixed models. *Stat. Sci.* **28**, 135–167 (2013).
24. West, B. T., Welch, K. B. & Galecki, A. T. *Linear mixed models: A practical guide using statistical software* (Chapman and Hall/CRC, London, 2014).
25. Josef Perktold, Skipper Seabold, Jonathan Taylor . statsmodels-developers (2009-2017). Available at: https://www.statsmodels.org/stable/index.html.
26. Lundervold, A. J., Vik, A. & Lundervold, A. Lateral ventricle volume trajectories predict response inhibition in older age—A longitudinal brain imaging and machine learning approach. *PLoS ONE* **14**, 1–19. https://doi.org/10.1371/journal.pone.0207967 (2019).
27. Lindstrom, M. J. & Bates, D. M. Nonlinear mixed effects models for repeated measures data. *Biometrics* **2**, 673–687 (1990).
28. Dietterich, T. G. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems* 1–15 (Springer, Berlin, 2000).
29. Saeys, Y., Abeel, T. & Van de Peer, Y. Robust feature selection using ensemble feature selection techniques. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 313–325 (Springer, Berlin, 2008).
30. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
31. Maass, A., Berron, D., Libby, L. A., Ranganath, C. & Düzel, E. Functional subregions of the human entorhinal cortex. *ELife* **4**, 2 (2015).
32. Schultz, H., Sommer, T. & Peters, J. The role of the human entorhinal cortex in a representational account of memory. *Front. Hum. Neurosci.* **9**, 628 (2015).
33. Cole, J. H. & Franke, K. Predicting age using neuroimaging: Innovative brain ageing biomarkers. *Trends Neurosci.* **40**, 681–690 (2017).
34. Edmonds, E. C. *et al.* Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. *Alzheimers Dement.* **11**, 415–424 (2015).
35. Edmonds, E. C. *et al.* Patterns of longitudinal cortical atrophy over 3 years in empirically derived mci subtypes. *Neurology* **2**, 2 (2020).
36. Arbabshirani, M. R., Plis, S., Sui, J. & Calhoun, V. D. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage* **145**, 137–165 (2017).
37. Idland, A.-V. *et al.* Biomarker profiling beyond amyloid and tau-CSF markers, hippocampal atrophy and memory change in cognitively unimpaired older adults. *Neurobiol. Aging* **2**, 2 (2020).
38. Nyberg, L. Neuroimaging in aging: Brain maintenance. *F1000Res* **6**, 1215 (2017).
39. Reuter-Lorenz, P. A. & Park, D. C. How does it STAC up? Revisiting the scaffolding theory of aging and cognition. *Neuropsychol. Rev.* **24**, 355–370 (2014).
40. Bellou, E. Age dependent effect of APOE and polygenic component of Alzheimer disease. *Neurobiol. Aging* **2**, 2 (2020).
41. van Loenhoud, A. C. *et al.* Cognitive reserve and clinical progression in Alzheimer disease: A paradoxical relationship. *Neurology* **93**, e334–e346 (2019).
42. Lundervold, A. S. & Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Z. Med. Phys.* **29**, 102–127 (2019).

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Ageing, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals

Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Data included in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Author contributions

S.A.M. was responsible for the design and analyses, interpretation of results, and drafting of the manuscript; A.J.L. and A.V. were responsible for interpretation of results and drafting of the manuscript, and A.S.L. contributed to the design and analyses, interpretation of results, and drafting the manuscript.

Funding

This work was supported by the Trond Mohn Research Foundation [Grant number BFS2018TMT07].

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.A.M.

Reprints and permissions information is available at www.nature.com/reprints.

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

PAPER C

FROM LONGITUDINAL MEASUREMENTS TO IMAGE CLASSIFICATION: APPLICATION TO LONGITUDINAL MRI IN ALZHEIMER'S DISEASE

Mofrad, Samaneh A., Bartsch, Hauke and Lundervold, Alexander S. From longitudinal measurements to image classification: Application to longitudinal MRI in Alzheimer's disease. *Under review.* April 2021.

From longitudinal measurements to image classification: application to longitudinal MRI in Alzheimer's disease

Samaneh A. Mofrad, Hauke Bartsch, Alexander S. Lundervold, and
for the Alzheimer's Disease Neuroimaging Initiative

***Abstract*—We propose a novel method of constructing representations of multiple one-dimensional longitudinal measurements as two-dimensional grey-scale images. This can be used to turn classification problems from longitudinal settings into simpler image classification problems, allowing for the application of newer deep learning methods on longitudinal measurements. To evaluate our approach, we apply it to an important and challenging task: the prediction of dementia from brain volume trajectories derived from longitudinal MRI. We construct an ensemble of convolutional neural network models to classify two groups of subjects: those diagnosed with mild cognitive impairment at all examinations (stable MCI) versus those starting out as MCI but later converting to Alzheimer's disease (converted AD). Models were trained on image representations derived from $N = 736$ subjects sourced from the ADNI database ($471/265$ sMCI/cAD). We obtained an accuracy of a resulting ensemble model of 76%, measured on an independent test set. Our approach is competitive (in terms of accuracy) with results reported in other machine learning approaches with similar classification on comparable tasks. This indicates that our approach can lead to useful representations of longitudinal data.**

***Index Terms*—Deep learning, Longitudinal data, Trajectories, Alzheimer's disease, Mild cognitive impairment, MRI.**

I. Introduction

Deep neural networks form the basis for a wide range of state-of-the-art medical image analysis tasks and has drawn a lot of interest over the past years [1], [2]. While deep learning techniques are behind successful applications in various fields, in many cases it is difficult to find an appropriate representation of the input data used to train deep learning models, that highlights the useful predictive features in the data.

Longitudinal data is one such case. Here measurements are taken repeatedly through time with multiple outcomes at each time point. Some of these difficulties are due to the inherent properties of longitudinal data, like inter-correlation between the set of observations of one subject [3] and the unbalanced observations for subjects [4].

Motivated by studies where time-series or speech recognition data were represented as images [5]–[7], we propose a pipeline for producing two-dimensional (2D) images from longitudinal data. This enables the use of well-studied techniques from deep learning for two-dimensional image classification.

First, we gathered all the data collected from each subject in a matrix so that one axis is associated with time points and the other to the corresponding values of those time points. Then, we scaled the columns' values separately to get a standard range for each variable. Next, we mapped each scaled matrix to a gray-scale image, so that the pixel intensity represents the matrix values (Fig. 1 illustrates the steps). The 2D images can then be used to train a deep neural network classifier.

To evaluate our proposed pipeline in a concrete setting, we used a longitudinal data source with a large number of subjects, which contains ascending, descending, and categorical data, where the number of time points and the length between them varies significantly. We used data from subjects diagnosed with various levels of dementia: Alzheimer's Disease (AD), which is a common irreversible neurodegenerative disorder characterized by a cognitive impairment that gradually worsens over time [8], [9], and Mild Cognitive Impairment (MCI), which is a transitional state from normal cognition to dementia [10]. We ran the experiment on two subgroups labeled as stable MCI (sMCI), who were diagnosed as MCI at all scans, and converged AD (cAD), who were diagnosed as MCI at the beginning but later developed AD.

After preparing 2D images for sMCI and cAD subjects, we investigated the effect of data augmentation techniques, model architectures, and hyper-parameter selection. We used the results from these investigations to construct an ensemble model that can classify conversion to AD versus stable MCI with an average accuracy of 76%. This is a competitive result when compared with other approaches, indicating the

Our work was supported by the Trond Mohn Research Foundation, grant number BFS2018TMT07

S.A.M and A.S.L are with Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Postbox 7030, 5020 Bergen, Norway (S.A.M email: sam@hvl.no)

S.A.M., H.B. and A.S.L. work with The Mohn Medical Imaging and Visualization Centre (MMIV), Department of Radiology, Haukeland University Hospital, Bergen, Norway

H.B. work with Department of Informatics, University of Bergen, Bergen, Norway

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

usefulness of the proposed pipeline also for other problems related to longitudinal measurement.

II. Methods

A. Data

Data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD, with an overall goal to validate biomarkers for use in clinical treatment trials for patients with AD. The study was approved by the Institutional Review Boards at each ADNI site (see full list here: http://adni.loni.usc.edu). Informed consent was obtained from all subjects prior to enrollment. All methods were carried out in accordance with relevant guidelines and regulation. The present study was approved with ADNI Publication Committee (ADNI DPC).

We constructed a longitudinal data set by selecting the ADNI subjects that had at least three MRI scans. Our data set consists of 1460 subjects (female/male: 642/818) with a total of 7421 MRI scans (see Table I for details). We considered five longitudinal labels based on the ADNI diagnoses, which were defined in our previous work [11]. If subjects were labeled as normal controls (CN) at all scans, we labeled as healthy controls (HC). The subjects who converted from CN to MCI during the study were labeled as converted MCI (cMCI). Subjects who were defined with MCI at all visits were labeled as stable MCI (sMCI), those converting from MCI to AD were labeled as converted AD (cAD), and those who were defined with AD at all visits were labeled as stable AD (sAD). This differentiates subjects into stable and progressive groups, and can be used to find the features associated with developing the conditions.

Group	Subjects	MRI	Gender (f/m)
HC	368	1916	187/181
cMCI	106	590	45/61
sMCI	471	2424	195/276
cAD	265	1532	104/161
sAD	250	959	111/139
	1460	7421	642/818

TABLE I: Longitudinal subjects and MR images: total number of subjects, MR images, and gender distribution within each of the five subgroups.

B. Image preparation

We used the measured volumes of all the regions in the brain that are extracted by Freesurfer [12] v.6.0 from the T1-weighted MR images. For each individual, we obtained such measurements at all time points, and thereby a two-dimensional matrix for each subject containing the volumes of brain regions at the time points. Further, to include their possible influence in our study, we added three more rows to the matrix: gender (male $= 0$, female $= 1$), the level of education (varies between 4 years and 20 years), and age of subject at MRI examinations (between 54 to 96). Therefore, for subject i ($i = 1, \dots, 1460$) we had a matrix x_i, so that $x_i \in \mathbf{R}^{m \times n_i}$, where $m = 125$ (number of ROIs + 3), and $3 \leq n_i \leq 11$ is the number of scans for subject i. The goal was then to construct a two-dimensional image for each subject based on its matrix.

We first selected 20% of subjects for the final test set at random, controlling for class labels (to have 20% of each label in the test set), gender (to have 20% of both male and female in the test set), and age (to have a similar range of age in both the training and test set). Then for the rest of the data (including subjects with less than three MRI scans), we assigned the ROIs and the three additional variables: age, gender, and education level as columns in a table, and inserted the extracted volumes from images into its rows (see Fig. 1a).

Next, we scaled the volumes to get them into a similar range. More specifically, volumes for regions of interest is measured in cubic millimeters and the way the calculation is done prevents negative values. However, in the regions of interest data, we observe a trend that exhibits a lower mean to be skewed toward few participants with very large volumes. In order to better utilize the limited resolution of the intensity values (8-12bit), we opted to perform a single sided winsorizing operation where the largest 2.5% of all the volumes values are replaced with less extreme values (Fig. 1b). We call these upper limits *robust max*. Then, we scaled each column of the table based on its minimum value and its robust max, $\frac{x_i \, - \, min}{robust \ max \, - \, min}$.

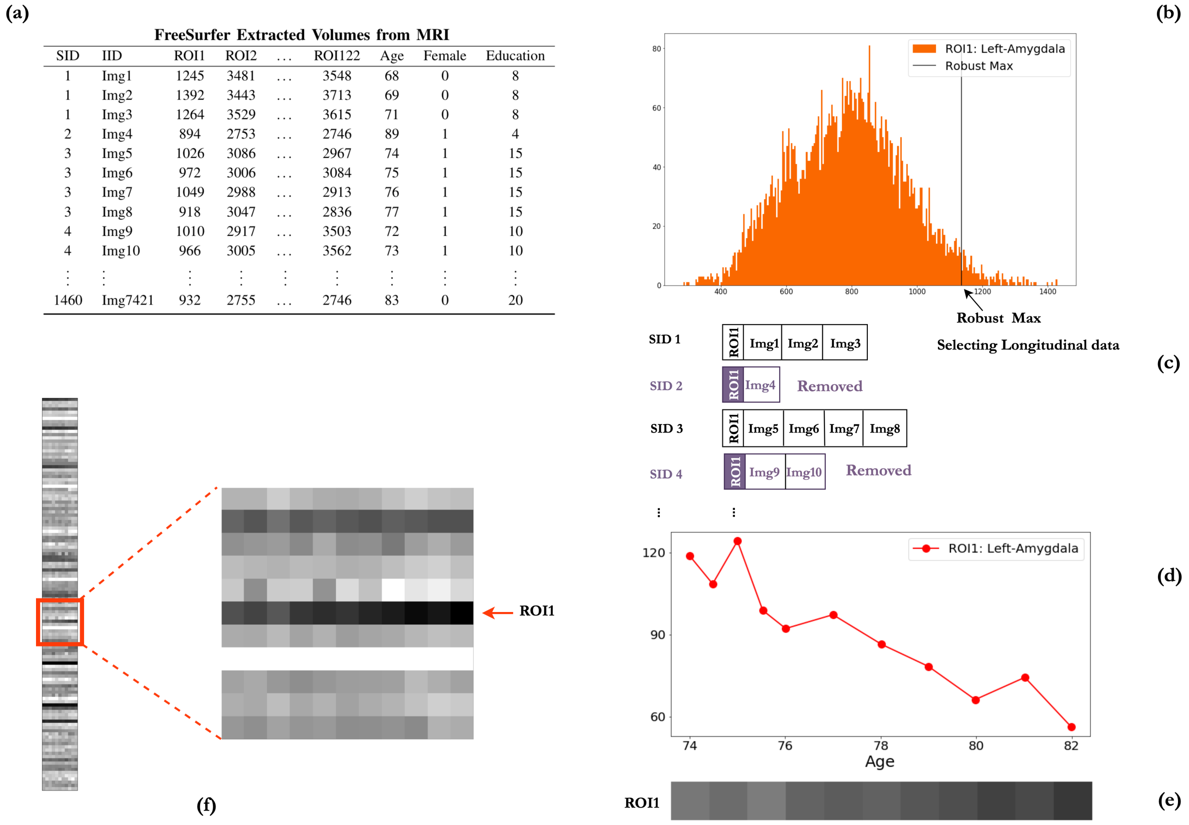
Note that the scaling of one subject is affected by all the other subjects in the table. To avoid data leakage, it was therefore important to separate a test set before scaling. The test set subjects, and potentially other new, previously unseen subjects, are scaled using the minimum and (robust) maximum computed using the training data.

After scaling the values in all columns, we selected the longitudinal subjects for which at least three MRI scans were available (Fig. 1c). Every subject has a volume-trajectory for each ROI (Fig. 1d) which we mapped to an image where the pixels' intensity in the image represents the ROI's scaled values at time points (see Fig. 1e, for one ROI). Then, we add the images of all ROIs on top of each other, plus the intensity images of age, education, and gender (Fig. 1h). This resulted in one image per participant, based on the volume extracted from the longitudinal MRI scans.

To determine if we can identify the differences between the prepared images in subgroups by their pixels' intensity we constructed the images in Fig. 2. For all subjects we linearly interpolated the values of ROIs to have the same image dimension for all subjects, and then we calculated the average of the matrices associated to all images in each subgroup. These average images (Fig. 2) highlight the differences in the intensities of more severe dementia (cAD and sAD) compared to healthy and MCI cases (HC, cMCI and sMCI).

C. Regularization techniques

During training of our models (Section II-D below), we used multiple regularization techniques. Both general explicit



FreeSurfer Extracted Volumes from MRI

SID	IID	ROI1	ROI2	…	ROI122	Age	Female	Education
1	Img1	1245	3481	…	3548	68	0	8
1	Img2	1392	3443	…	3713	69	0	8
1	Img3	1264	3529	…	3615	71	0	8
2	Img4	894	2753	…	2746	89	1	4
3	Img5	1026	3086	…	2967	74	1	15
3	Img6	972	3006	…	3084	75	1	15
3	Img7	1049	2988	…	2913	76	1	15
3	Img8	918	3047	…	2836	77	1	15
4	Img9	1010	2917	…	3503	72	1	10
4	Img10	966	3005	…	3562	73	1	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1460	Img7421	932	2755	…	2746	83	0	20

Fig. 1: Here we illustrated an example of preparing images from brain regions volumes extracted from MR images. The left part of the figure is for all ROIs, where on the right side, we explain a specific ROI. Note that we should first detach the test set. **a)** We assign the ROIs to the columns of a table where each row corresponds to the volumes of ROIs for one image. The number of MR images varies from one ID to another. **b)** For each ROI, we find a robust max and replace the upper outliers with this value. Then, we scale the volumes in the column by the min and robust max between 0 and 255. **c)** Next, we select the longitudinal subjects which have at least three images. The graph in **(d)** shows the left-Amygdala scaled volumes versus age. **e)** This graph maps to a gray-scale image so that pixel intensity represents the changes in the values. **f)** Finally, we attach the gray images of all ROIs on top of each other to get an image for each subject.

techniques such as dropout, batch normalization, and weight decay, and data augmentation tailored to our specific data set, as described here.

Aiming to balance the class sizes and to use a source of variance in our data set to boost our models' generalization ability, we augmented the data set by adding Gaussian noise to the existent images. We presume that the obtained volumes for ROIs contain noise (confer the instability in trajectory graph in Fig. 1d), which likely are related to the physical and biological situation during scanning, uncertainly concerning the quality of T1 weighted images, and our chosen segmentation tool (FreeSurfer). To estimate how the variability in the volumes of the ROIs produced by repeated scans in a short time affects our constructed 2D images, we identified 14 subjects in ADNI who had at least two MRI scans within a month or less. It's natural to assume that the volumes of one's brain regions change very little over one month, but comparing the extracted volumes of ROIs for these two repeated MRI scans showed differences in volume (from $\pm 5.3\ mm^3$ for Left-vessel to $\pm 5563.2\ mm^3$ for Cerebral White Matter volume). For each ROI, we averaged 14 standard deviations, measured separately for two collected volumes of each subject, and called it σ_{roi}. Then, we added Gaussian noise with zero mean and the measured standard deviation for each ROI ($P_{roi}(\mu = 0, \sigma_{roi})$) to the training set until we collected 600 subjects for each class. Then we incorporated the noisy data in one table. Afterward, we normalized the new table by using the min and robust max saved for each ROI (Fig. 1b) and then prepared images based on the noisy versions of existing subjects by following the steps in Fig. 1(c) to 1(f).

D. Model selection

Since model performance is typically very sensitive to hyper-parameter tuning, we performed an extensive search over a wide set of hyper-parameters. To find the optimized values for learning rate, weight decay, dropout, and the CNN structures, we selected 10 different training-validation sets (hereafter 10 folds). For each label, we randomly selected 18% of the training set in order to keep the same percent of gender and the same range of age in both training and validation sets

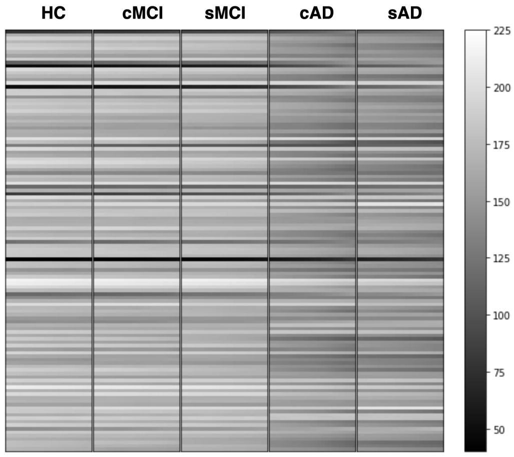


Fig. 2: The average image for all subjects in five longitudinal subgroups after normalization and interpolation.

when possible. Note that we put the test set aside before this step. A grid search over model architectures and these hyper-parameters was conducted by varying the following:

- CNN model: we considered ResNet18 and ResNet34, 18-layer and 34-layer residual networks [13], as implemented in the Torchvision library [14].
- Probability of dropouts on the hidden layers (ps): we passed eight values for ps: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7.
- Weight decay (wd): for wd we passed four values: 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}.
- Maximum learning rate (max-lr): these five values were tested for max-lr: 3×10^{-2}, 3×10^{-3}, 3×10^{-4}, 1.5×10^{-4} and 1×10^{-4}.

To construct and train our binary CNN classifiers we used fastai [15] version 1.0.61, a deep learning library based on PyTorch. Instead of training all layers with a constant learning rate or decreasing the learning rate with a fixed or exponential value, we applied the cyclical learning rates method [16] as implemented in fastai, which varies the learning rate cyclically between a reasonable set of minimum and maximum boundaries [15]. The batch size was set to eight for all models.

We also compared the performance of the ResNet models with and without pretraining, using the pretrained ResNet18 and ResNet34 models available in PyTorch, fine-tuning them on our data using fastai. Further, we investigated whether batch normalization had a significant effect on the performance of models.

During model selection, we monitored the training and validation loss, error rates, accuracy, precision, recall, and F_1 score on the validation set. For each combination of parameters, we estimated the optimal number of epochs by finding the epochs associated with the smallest validation loss separately for all 10 folds and computing their average.

After the grid search, we selected the top performing models in terms of accuracies over the 10 validation sets, and we got our final results by ensembling these models using both soft and hard voting strategies. In hard voting the ensemble predicts the majority vote among the individual models, while soft voting is based on averaging the class probabilities of the models.

There are several sources of randomness in the PyTorch CNN models, leading to slightly different results every time a model is used. To limit the effect of such randomness, we fixed the random seed in both Numpy and Pytorch. Further, to also reduce the effect of other sources of randomness, such as dropout layers, we trained the ensemble models 20 times with both hard and soft voting and reported the average and standard deviation for the final results.

Finally, to investigate whether spatial relationships reflected in the ordering of the various ROI measurements influence the model, we randomly shuffled the order of ROIs in the images ten times. We then applied the same ensemble model to the identical pair of training and test sets for these ten different image sets.

III. Results

Our results are based on the steps described in Section II-D applied to two subgroups of subjects, sMCI and cAD (see Fig. 2).

A. Model selection

Fig. 3a shows the similarity in the performance (accuracy) of the models with and without batch normalization (A and B, respectively), and also the significant decline in the performance when using the pretrained ResNet18/ResNet34 models with batch normalization (C). As our images are quite different from the images used for pre-training, the low performance of this model is perhaps to be expected. During the grid-search, we saved the number of epochs associated with the average of lower validation loss for 10 validation sets. Fig. 3b shows the similarity in the number of epochs for models with and without batch normalization (A and B) while the number of epochs for the case with transfer learning (C) is significantly smaller. Thus, while transfer learning speeds up the training step, the accuracy is not as high as cases A and B. We chose to use model A in the following, i.e. without pretraining and with batch normalization.

To explore the value of our data augmentation approach, we repeated the experiment once again by duplicating the existing images instead of adding Gaussian noise. Fig. 4 compares the performance of the models when augmentation is according to the Gaussian noise ($\mu = 0,\ \sigma_{roi}$) with the case of duplicating the available images. There is an insignificant difference between the accuracies (Fig. 4a) and the number of epochs associated with the average of lowest validation loss (Fig. 4b).

B. Ensemble model

Based on the results of the previous section we selected the top 19 models (see Table II) to investigate whether constructing an ensemble model improves the results compared to the

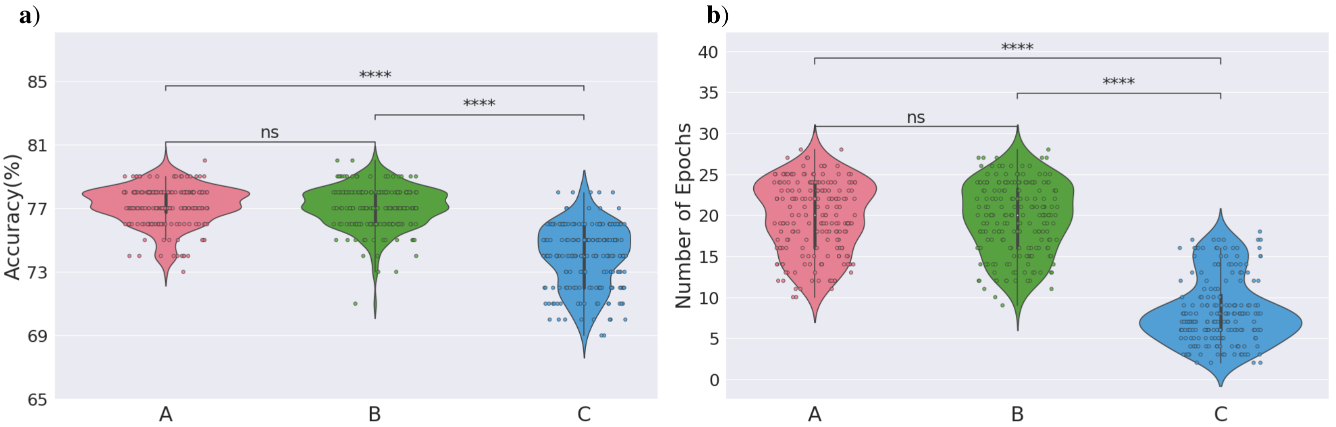


Fig. 3: Comparing the effect of batch-normalization and transfer learning during grid-search; A: Batch-normalization and non-pretraining ResNet18/ResNet34) models, B: without batch-normalization and non-pretraining, C: batch-normalization and pretraining. The average accuracies **a)** and epochs associated with the lower validation loss **b)** over validation sets show the similarity between A and B, while in C are reduced significantly. P-value; ns: $p > 0.05$ and ****: $p < 0.0001$.

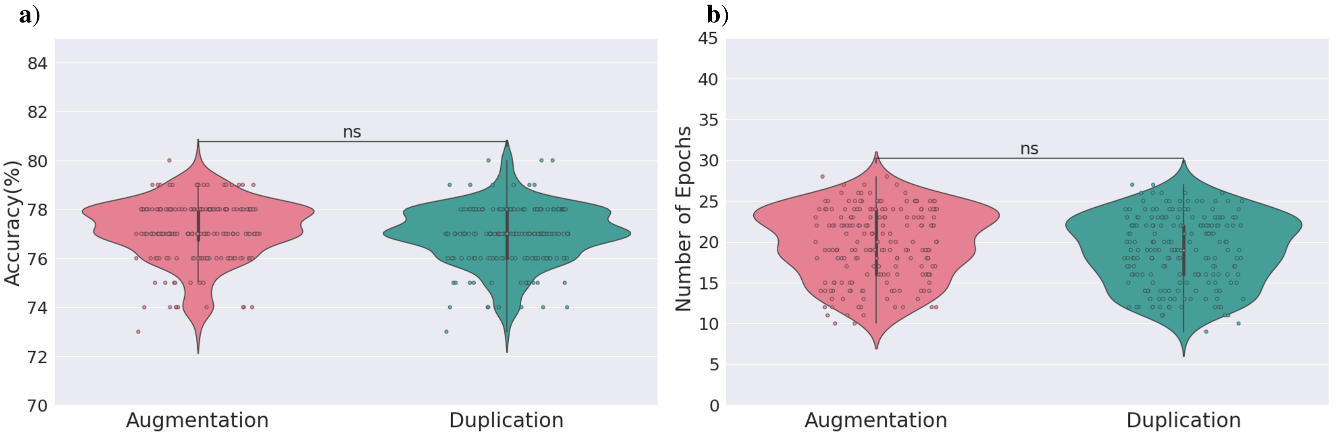


Fig. 4: The comparison between grid-search over data with Gaussian noise and duplicating the existing images shows an insignificant difference between the accuracy **a)** and the number of epochs associated with the lowest validation loss **b)**. P-value; ns: $p > 0.05$.

Models	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
ResNet	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	34	34	34
epoch	13	15	21	14	15	14	13	16	17	16	16	19	16	17	17	20	16	13	17
ps	1e-1	2e-1	2e-1	2e-1	4e-1	4e-1	4e-1	5e-1	5e-1	5e-1	5e-1	6e-1	6e-1	6e-1	6e-1	7e-1	3e-1	3e-1	5e-1
wd	1e-4	1e-1	1e-2	1e-3	1e-1	1e-1	1e-4	1e-1	1e-2	1e-3	1e-1	1e-2	1e-1	1e-2	1e-3	1e-2	1e-1	1e-3	1e-4
max-lr	3e-4	3e-4	3e-3	3e-4	3e-4	1.5e-4	1.5e-4	3e-4	3e-4	3e-4	1e-4	3e-4	1e-4	1.5e-4	1e-4	3e-4	3e-4	1.5e-4	3e-4

TABLE II: 19 selected models based on: network architectures (ResNet18 and ResNet34), hyper-parameters (ps: probability of dropout layers, wd: weight decay, max-lr: maximum learning rate), and number of epochs.

individual models in terms of accuracy and robustness. Fig. 5a shows the average of receiver operating characteristic (ROC) curves of all models for a specific validation set. The mean and standard deviation of areas under the ROC curves (ROC AUC) for each validation set was computed. These curves (Fig. 5a) highlight the differences in model performance over the validation sets.

For each model, we averaged the ROC curves across ten validation sets (Fig. 5b). Based on the curves in Fig. 5b, the differences between the ROC AUC mean and standard deviation of the 19 models are insignificant (between 0.79 ± 0.04 and 0.81 ± 0.04). Therefore, while the models differ in their performance on single validation sets (Fig. 5a), their averages over all the validation sets are quite close (Fig. 5b).

Further, the mean ROC AUCs for the ensemble model, applied separately to each validation set, is plotted in black

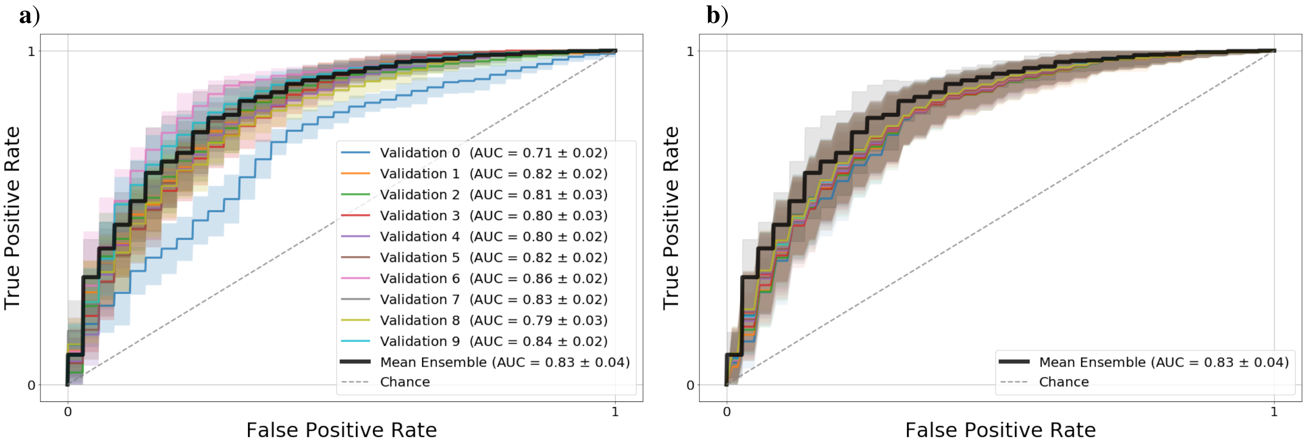


Fig. 5: **a)** Averaged ROC curve and standard deviation of AUC based on a specific validation set (color) for top 19 selected models. **b)** Averaged ROC curve and standard deviation of AUC based on a specific model (color) for all validation sets. While the models have different performance on each specific validation set (**a**), the average of performance on all the validation sets have similarities (**b**). The average ROC for the ensemble model is in black.

in Fig. 5a and b. The ROC AUC mean value and standard deviation for the ensemble model are 0.83 ± 0.04, which is higher than the top individual ROC AUC on seven validation sets (Fig. 5a), and higher than all ROC AUC for individual models when averaged over ten validation set (Fig. 5b).

C. Classification of sMCI versus cAD

The final evaluation of our models was conducted by training the 19 selected models on the combined training and validation sets and forming two ensemble models, based on soft and hard voting. Their performance on the separate test set was then computed. To reduce the effect of randomness, we repeated the computations 20 times and averaged the results. Fig. 6 shows the final results for the soft and hard voting ensembles. The averages of accuracy, weighted precision, recall, and F_1score for hard voting are as follows: 75.9%, 77.1%, 75.9%, and 76.1% respectively (Fig. 6a). The averages for soft voting are as follows: 76.3%, 77.5%, 76.3%, and 76.6% respectively (Fig. 6b). Fig. 6c represents the accuracies of 20 runs for both hard and soft voting.

Finally, we randomly shuffled the order of ROIs in the images 10 times, and evaluated the same 19 models on these images. The accuracies ranged from 72% to 80%, with an average of 75.5%.

IV. Discussion

We proposed a method to represent the information of longitudinal metadata as two-dimensional images, enabling the construction of image-based machine learning classifiers.

We evaluated the method in an experiment based on the ADNI data set, aiming to classify stable MCI versus converged AD subjects using convolutional neural network models. We achieved higher-than-chance results, with an average accuracy of 76%. Our results show that our proposed method is competitive with other CNN-based approaches in the literature [17], where the reported accuracies range from 62% to 83% in similar setups, based on varying machine learning methods and multi-modal data sources ([11], [17]–[21]).

Our study has some limitations related to the chosen source of data. As shown in Fig. 1, the trajectories of brain volume exhibit some instability. This instability affects pixel intensity, and therefore makes the classification more challenging. Another limitation of this study is the difficulty in diagnosing MCI and AD. Some studies have shown the establishment of AD in the brain years before cognitive impairments appear in the behavioral functionality of the brain [22]. In the data from ADNI, the time span between visits for subjects are half a year on average, and we potentially have some sMCI that are almost AD. A possibility to make the model more robust is to drop one or two last visits of the sMCI subjects since they may be showing AD symptoms in less than six months.

Our results indicate that the proposed approach to longitudinal data analysis can be suitable as a supplementary analysis method next to more established statistical and machine learning analysis streams. Further assessment requires applying the method to multiple different data sources with more diversity in the form of data, such as ascending, descending, categorical, cyclical, or heterogeneous trends over time. An interesting area to explore would be time-series of resting-state functional MRI or longitudinal data of other progressive diseases such as schizophrenia or Parkinson's disease, which has different patterns of changes in brain volumes [23]–[25].

We hope that this integration of newer machine learning methods will create additional avenues for researchers to work on longitudinal data.

V. Acknowledgment

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Ageing, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie,

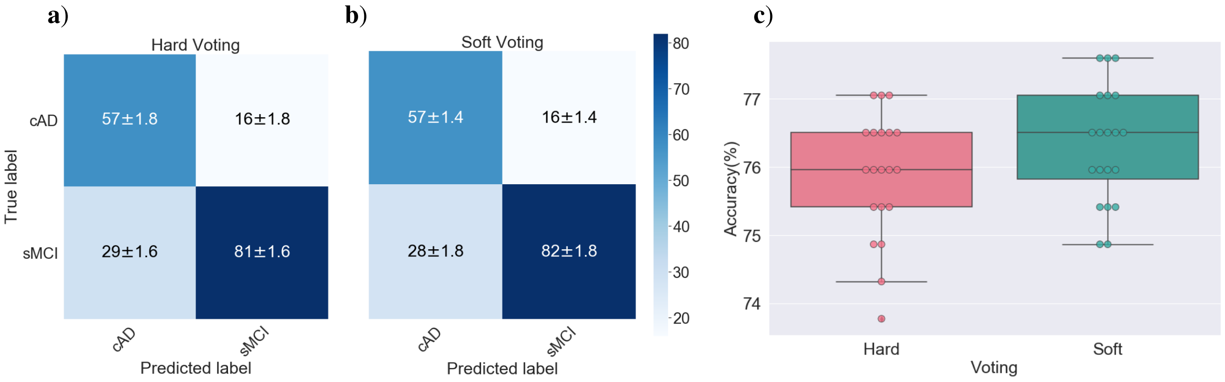


Fig. 6: **a)** The confusion matrix after hard voting over 19 models obtained the average accuracy, precision, recall, and F_1score of 76%,77%, 76%, and 76% respectively. **b)** Soft voting over 19 models obtained accuracy, precision, recall, and F_1score of 76.%, 78%, 76%, and 77% respectively. **c)** Boxplot for accuracy of 20 runs of ensemble models.

Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[2] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[3] P. Diggle, P. J. Diggle, P. Heagerty, K.-Y. Liang, P. J. Heagerty, S. Zeger *et al.*, *Analysis of longitudinal data*. Oxford University Press, 2002.

[4] D. A. Grimes and K. F. Schulz, "Cohort studies: marching towards outcomes," *The Lancet*, vol. 359, no. 9303, pp. 341–345, 2002.

[5] M. Kalash, M. Rochan, N. Mohammed, N. D. Bruce, Y. Wang, and F. Iqbal, "Malware classification with deep convolutional neural networks," in *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*. IEEE, 2018, pp. 1–5.

[6] D. Shulga, V. Silber-Varod, D. Benson-Karai, O. Levi, E. Vashdi, and A. Lerner, "Toward Explainable Automatic Classification of Children's Speech Disorders," in *International Conference on Speech and Computer*. Springer, 2020, pp. 509–519.

[7] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K.-K. R. Choo, "Imaging and fusing time series for wearable sensor-based human activity recognition," *Information Fusion*, vol. 53, pp. 80–87, 2020.

[8] A. P. Association, *Diagnostic and statistical manual of mental disorders (DSM-5)*. Pilgrim Press, Washington, 2013.

[9] D. C. Park and P. Reuter-Lorenz, "The adaptive brain: aging and neurocognitive scaffolding," *Annual review of psychology*, vol. 60, pp. 173–196, 2009.

[10] Y. E. Geda, "Mild cognitive impairment in older adults," *Current psychiatry reports*, vol. 14, no. 4, pp. 320–327, 2012.

[11] S. A. Mofrad, A. Lundervold, and A. S. Lundervold, "A predictive framework based on brain volume trajectories enabling early detection of Alzheimer's disease," 2020, submitted.

[12] B. Fischl, "FreeSurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] T. Contributors, "Torchvision. models," 2018.

[15] J. Howard and S. Gugger, "Fastai: A layered API for deep learning," *Information*, vol. 11, no. 2, p. 108, 2020.

[16] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 464–472.

[17] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot *et al.*, "Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation," *Medical Image Analysis*, p. 101694, 2020.

[18] R. Cui, M. Liu, A. D. N. Initiative *et al.*, "Rnn-based longitudinal analysis for diagnosis of alzheimer's disease," *Computerized Medical Imaging and Graphics*, vol. 73, pp. 1–10, 2019.

[19] Y. Shmulev, M. Belyaev, A. D. N. Initiative *et al.*, "Predicting conversion of mild cognitive impairments to Alzheimer's disease and exploring impact of neuroimaging," in *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities*. Springer, 2018, pp. 83–91.

[20] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's Disease diagnosis using structural MRI," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[21] S. A. Mofrad, A. J. Lundervold, A. Vik, and A. S. Lundervold, "Cognitive and MRI trajectories for prediction of Alzheimer's disease," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[22] C. R. Jack Jr and D. M. Holtzman, "Biomarker modeling of Alzheimer's disease," *Neuron*, vol. 80, no. 6, pp. 1347–1358, 2013.

[23] R. Yilmaz, F. Hopfner, T. van Eimeren, and D. Berg, "Biomarkers of Parkinson's disease: 20 years later," *Journal of Neural Transmission*, vol. 126, no. 7, pp. 803–813, 2019.

[24] X. Fei, Y. Dong, H. An, Q. Zhang, Y. Zhang, and J. Shi, "Impact of region of interest size on transcranial sonography based computer-aided diagnosis for Parkinson's disease," *Mathematical Biosciences and Engineering*, vol. 16, no. 5, pp. 5640–5651, 2019.

[25] D. Rodrigues-Amorim, T. Rivera-Baltanás, M. López, C. Spuch, J. M. Olivares, and R. C. Agís-Balboa, “Schizophrenia: a review of potential biomarkers,” *Journal of psychiatric research*, vol. 93, pp. 37–49, 2017.

PAPER D

ON NEURAL ASSOCIATIVE MEMORY STRUCTURES: STORAGE AND RETRIEVAL OF SEQUENCES IN A CHAIN OF TOURNAMENTS

Mofrad, Asieh A.*, Mofrad, Samaneh A.*, Yazidi, Anis and Parker, Matthew G. 6 April, 2021. On neural associative memory structures: Storage and retrieval of sequences in a chain of tournaments. *Neural Computation*. MIT Press. *: equal contribution.

On Neural Associative Memory Structures: Storage and Retrieval of Sequences in a Chain of Tournaments

Asieh Abolpour Mofrad[1, *]
Samaneh Abolpour Mofrad[2, 3, *]
Anis Yazidi[4, 5]
Matthew Geoffrey Parker[1]

[1] The Selmer Center, Dept. of Informatics, University of Bergen, Bergen, Norway.

[2] Dept. of Computer Science, Electrical Engineering, and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway.

[3] Mohn Medical Imaging and Visualization Center, Haukeland University Hospital, Bergen, Norway.

[4] Dept. of Computer Science, OsloMet - Oslo Metropolitan University, Oslo, Norway.

[5] Dept. of Plastic and Reconstructive Surgery, Oslo University Hospital, Oslo, Norway.

[*] These authors contributed equally.

Keywords: Auto-associative memories, clique-based neural networks, tournament-based neural networks, sequence storage

Abstract

Associative memories enjoy many interesting properties in terms of error correction capabilities, robustness to noise, storage capacity and retrieval performance and their usage spans over a large set of applications. In this article, we investigate and extend Tournament-Based Neural Networks, originally proposed by Jiang et al. (2016), which is a novel sequence storage associative memory architecture with high memory efficiency and accurate sequence retrieval. We propose a more general method for learning the sequences which we call Feedback Tournament-Based Neural Networks. The retrieval process is also extended to both directions: forward and backward, i.e. any large-enough segment of a sequence can produce the whole sequence. Furthermore, two retrieval algorithms, Cache-Winner and Explore-Winner are introduced to increase the retrieval performance. Through simulation results, we shed light on the strengths and weaknesses of each algorithm.

1 Introduction

Neural associative memory is a type of neural networks which is capable of memorizing (learning) a set of patterns and retrieving them from their corresponding noisy or incomplete versions. The term *association* refers to the linkage of two or more pieces of information. Hopfield neural network (Hopfield, 1982) was among the first designed artificial neural network with auto-associative memories which is able to retrieve information given only some partial clues as well as reconstruct perturbed patterns. Hopfield neural networks have some drawbacks such as being biologically implausible, due to the fully connected structure, low efficiency and spurious memories (see, e.g., Hoffmann, 2019, and references therein). To improve Hopfield network many variants of it have been proposed in the literature (see, e.g. Maurer et al., 2005; Berrou & Gripon, 2010; Krotov & Hopfield, 2016; Kim et al., 2017). Due to the *sparse coding* in the brain (for sparse coding see, e.g, Olshausen & Field, 2004; Rinkus, 2010), sparse associative memories are considered more biologically plausible models (Gripon et al., 2016; Hoffmann, 2019).

Gripon & Berrou (2011) proposed novel sparse neuro-inspired associative memories that organize neurons into clusters and memorize patterns using the concept of cliques (see also, Hopfield, 2008, for another clique-based network model of associative memory). This model, also referred to as GB model or Clustered Cliques Networks (CCNs), has fundament in information theory (Gripon & Berrou, 2012) and bears similarity to the Willshaw-type model (Willshaw et al., 1969) where sparse patterns and binary connections are considered. These models have been further developed in the literature (e.g. Aliabadi et al., 2014; Boguslawski et al., 2014; Jarollahi et al., 2014, 2015; Jiang et al., 2015, 2016; Mofrad et al., 2015, 2016; Mofrad & Parker, 2017; Berrou & Kim-Dufor, 2018), and used in many applications, such as solving feature correspondence problems (Aboudib et al., 2016), devising low-power content-addressable memory (Jarollahi et al., 2015), oriented edge detection in image (Danilo et al., 2015), image classification with Convolutional Neural Networks (Hacene et al., 2019), finding all matches of a probe in a database (Hacene et al., 2017), to mention a few. Furthermore, they were implemented on a general purpose graphical processing unit (GPU) (Yao et al., 2014), in 65-nm CMOS (Larras et al., 2018), and in distributed smart sensors architectures (Larras & Frappé, 2020). Therefore, CCN models can be referred to as an important brain-inspired memory system (Berrou et al., 2014) that became a basis for a wide range of research in associative memory models.

Learning and retrieval of temporal sequences in neural networks is a fundamental property of human intelligence which is studied through different approaches (see, e.g., Brea et al., 2011; Hawkins et al., 2009; Maurer et al., 2005; Jiang et al., 2016). Tournament-based Neural Network (TNN) (Jiang et al., 2016) is an extension of the clique-based approach to associative memories which have oriented connections, and therefore the ability to store sequential information (see also, Marques et al., 2017, for an implementation on the GPU). The novel structure of TNN is not only a sequence storage with high memory efficiency, but also a more compatible model with the neuronal signal propagation in the brain via oriented connections (see also Hawkins et al., 2009; Hawkins & Ahmad, 2016, for biologically plausible memory sequence structures).

In this paper, we improve the TNN architecture by proposing a more general structure, named Feedback TNN, as well as more accurate retrieval algorithms. The original TNN can be considered as a special case of Feedback TNN, with zero feedback connections. For retrieval, obviously, a less number of random selections during retrieval results into less component and sequence error at the end. The Cache-Winner retrieval revisits and changes some previous randomly selected components, in case an error is detected during retrieval. On the other hand, Explore-Winner reduces the randomness in decisions by considering the consequences of each decision. The idea behind the Cache-Winner technique can be illustrated in simple terms by drawing analogy with human decision making: imagine a person who makes a decision fast and then, if he realizes a mistake, tries to resolve it by manipulation of past decisions. On the other hand, Explore-Winner has the analogy with a rather careful decision-maker who investigates the consequences of all possible decisions at the time and then makes the best possible decision. In terms of achieving accurate sequence retrieval, both proposed retrieval techniques are superior to the Winner, which literally makes a random decision in the case of equal chance situations, and continues without further actions even when realizing a mistake later.

It is also known that the brain is able to follow the previously stored sequences, from any given point forward, and somewhat, also backwards (see, e.g. Hawkins & Blakeslee, 2007). The other contribution of this paper is introducing Feedback-Backward retrieval method which makes our model more biologically plausible. Using Feedback-Backward retrieval, the model gains the capability of retrieval of the whole sequence, given a sub-sequence, no matter its location. The Feedback-Backward retrieval is more compatible with the Feedback TNN, but works well with the original TNN as shown in the results. Backward retrieval, therefore, adds more capabilities to these types of sequence storage structures, and makes them more similar to brain functioning.

The paper is organized as follows: in section 2 we briefly survey the CCN and TNN structures. In section 3, different learning and retrieval algorithms are explained. The simulation results are provided in section 4, and afterwards, in section 5, discussion and concluding remarks are presented.

2 Background

In this section, first the clustered clique-based neural network structure is described in section 2.1. These types of networks are able to store and retrieve the fixed length patterns. Next, in section 2.2, tournament-based neural networks which have the ability to store and retrieve sequences is surveyed.

2.1 Clustered Clique Networks (CCNs)

In Clustered Clique Networks (CCNs) the way the neurons are organized within clusters, and the sparsity of the encoding used for storing patterns in cliques, result into large storage diversity, i.e. number of storable patterns, high capacity, i.e. the amount of storable information, and strong robustness against erasures and errors (Gripon & Berrou, 2011; Jarollahi et al., 2015; Gripon et al., 2016).

Formally, the structure of CCNs consists of n neurons divided into c clusters with possibility of different sizes. The input patterns are formed from a pre-defined alphabet $\mathcal{A}$ where the number of neurons in each cluster matches the size of used alphabet $|\mathcal{A}|$. For simplicity all clusters are considered to have the same number of neurons, say $l = n/c$, and therefore the same alphabet size $|\mathcal{A}| = l$. The j^{th} neuron in the i^{th} cluster is denoted by n_{ij} and it has an associated value, $v(n_{ij})$, equals one if it is activated, and zero otherwise; where $1 \leq i \leq c$ and $1 \leq j \leq l$. Let $\mathcal{P}$ be the set of patterns to be stored where pattern $p \in \mathcal{P}$ contains c sub-patterns, i.e. $p = p_1 p_2 \cdots p_c$; for $p_i \in \mathcal{A}$.

The learning process starts by assigning a unique set of neurons -one per cluster- to each $p \in \mathcal{P}$:

$$p = p_1 p_2 \cdots p_c \rightarrow (f(p_1), f(p_2), \cdots, f(p_c))$$
$$\text{where} \quad f : \{p_i\} \rightarrow \{n_{ij} | 1 \leq j \leq l\}.$$

Learning proceeds by activation of the selected neurons, i.e. $v(n_{ij}) = 1$, and forming a clique by connecting the selected c active neurons to each other through binary edges. As a result, the learning process generates a set of binary edges

$$\mathcal{W} = \{\omega_{(ij)(i'j')} | \text{ if } i \neq i' \text{ and } \exists\, p \in \mathcal{P} \text{ s.t. } f(p_i) = n_{ij} \text{ and } f(p_{i'}) = n_{i'j'}\},$$

where $\omega_{(ij)(i'j')}$ is an edge between n_{ij} and $n_{i'j'}$.
The edge $\omega_{(ij)(i'j')}$ belongs to $\mathcal{W}$ independently from the number of patterns that use both n_{ij} and $n_{i'j'}$ neurons, but only if there exists such a pattern. Figure 1 illustrates the storing process in clique-based networks.

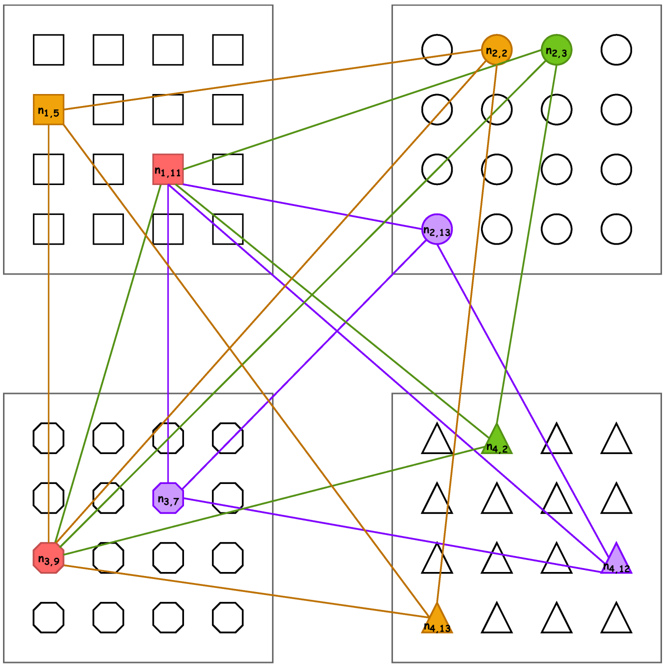


Figure 1: The learning process of three patterns, in a network with $c = 4$ clusters and $l = 16$ neurons per cluster. Node $n_{i,j}$ refers to the j^{th} neuron in the i^{th} cluster. Each clique represents one of the three $(4, 1, 8, 12)$, $(10, 2, 8, 1)$, and $(10, 12, 6, 11)$ patterns with yellow, green, and purple respectively. Coloured nodes refer to the activation of neurons for at least one pattern. The red nodes, $n_{1,11}$ and $n_{3,9}$, belong to two patterns. Note that it is not possible to retrieve the patterns by finding a unique clique using only one of these red nodes.

The recall or retrieval phase of a possibly distorted version of a learnt pattern, $\hat{p}$, is based on finding the closest match from $\mathcal{P}$. Depending on the type of distortion, various retrieval methods might be used (see, Aboudib et al., 2014), however, in general the recall procedure consists of local and global phases. The local phase aims to find the most probable neurons in different clusters, using information from $\hat{p}$ or incoming connections from previously activated neurons, and activate them, i.e. $v(n_{ij}) = 1$. The global phase is to recall the established edges in $\mathcal{W}$ that have an end in activated neurons. This procedure alternate between global and local retrieval to gradually complete the clique and therefore the pattern.

It is noteworthy that other sparse structures were presented by Aliabadi et al. (2014), according to which, $c \ll \chi$ where $\chi = n/l$ denotes the number of clusters and c was used to denote a smaller set of clusters for which a sparse pattern is mapped into. Retrieval, in this case, would be more complicated and various scenarios could be considered (see, e.g., Aboudib et al., 2014; Jiang, 2014). For instance, the winner-take-all rule activates neurons with the highest activity (or maximum score), whilst Losers Kicked-Out rule (LsKO) eliminates active neurons with less activity using a threshold filter (see Jiang, 2014, for details).

2.2 Tournament-Based Neural Network (TNN)

An extension of the CCNs (Jiang et al., 2016) is proposed by using directed edges between clusters in such a way that the network can store sequential information in a tournament-based[1] neural network. In a chain of tournaments of order c and degree r, denoted by $\mathcal{T}_r(c)$, each node is directed clockwise to its r consecutive neighbors; see Figure 2 with $c = 8$ and $r = 3$ for a sample chain of tournaments. A TNN can then be seen as a concatenation of tournaments of size $r + 1$.

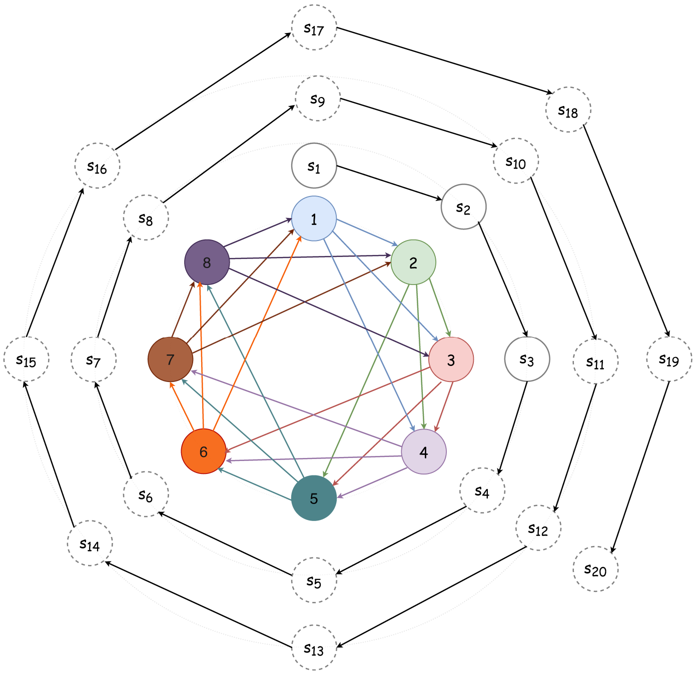


Figure 2: An illustration of a chain of tournaments, $\mathcal{T}_3(8)$, for storing sequences of length 20. The eight clusters are represented by colored circles, and each arrow represents a set of possible connections between nodes within the clusters. The clusters construct eight tournaments of size $r + 1 = 4$. For instance, clusters that have been shown with $1, 2, 3, 4$ make one tournament starting from cluster 1, and clusters labeled with $7, 8, 1, 2$ involve in another tournament starting from cluster 7. A sequence of length 20 and the assigned clusters for each component s_i are represented around the network. Given the first r components (s_1, s_2, s_3) with solid circles, the retrieval algorithm could retrieve the rest sequentially using the tournament connections. This figure is based on (Jiang et al., 2016, Fig. 5).

In order to store a set of sequences, $\mathcal{S}$, in a chain of tournaments, we suppose that each sequence $s \in \mathcal{S}$ contains L component, i.e. $s = s_1 s_2 \cdots s_L$; for $s_t \in \mathcal{A}$, $t = 1, 2, \ldots, L$, and $|\mathcal{A}| = l$.

By labeling clusters from 1 to c, the learning process could be explained as follows. First a unique sequence of neurons must be assigned to each $s \in \mathcal{S}$ by using function

[1] In graph theory, by assigning direction to all edges of a complete graph, a tournament can be achieved.

$f = (f_1, \cdots, f_c)$, where $f_i,\ i = (t-1 \bmod c)+1$, maps a component s_t, to a unique neuron n_{ij} in cluster i:

$$f_i : \{s_t\} \to \{n_{ij} | 1 \leq j \leq l, \},\ 1 \leq i \leq c,$$

therefore,

$$f(s) = (f_1(s_1), f_2(s_2), \cdots, f_c(s_c), \cdots, f_{(L-1 \bmod c)+1}(s_L))$$

Learning continues by connecting neuron n_{ij} to neuron $n_{i'j'}$ at passage π as follows

$$n_{ij} \to n_{i'j'},\ \text{ if: } \begin{cases} f_i(s_{(i+(\pi-1)c)}) = n_{ij} \\ f_{i'}(s_{i'+(\pi-1)c}) = n_{i'j'} \end{cases} \text{ and, } 1 \leq \delta_i(i') \leq r \tag{1}$$

where $\delta_i(i') = (i'-i) \bmod c$, and $1 \leq \pi \leq \lfloor \frac{L}{c} \rfloor$.

In general, for $s \in \mathcal{S}$, if the above conditions are satisfied for a given π such that $n_{ij} \to n_{i'j'}$, we set $N_{s,\pi}(n_{ij}, n_{i'j'}) = 1$, which means that n_{ij} is connected to $n_{i'j'}$, in sequence s, otherwise we set $N_{s,\pi}(n_{ij}, n_{i'j'}) = 0$. In Figure 2, s_2 is connected to s_3 in passage $\pi = 1$, but not to s_{11} (in passage $\pi = 2$), and s_{19} (in passage $\pi = 3$) in the same sequence s, for instance. So the neighboring connections are defined based on both s and π values.

At the end of learning or storing process, the network has the following connections:

$$\mathcal{W} = \{\omega_{(ij)(i'j')} | \text{ if } \exists\ s \in \mathcal{S},\ \text{ and } \exists\ \pi \in [1 : \lfloor \frac{L}{c} \rfloor] \text{ s.t. } N_{s,\pi}(n_{ij}, n_{i'j'}) = 1\} \tag{2}$$

where $\omega_{(ij)(i'j')}$ is a directed edge from n_{ij} to $n_{i'j'}$ and $1 \leq i, i' \leq c,\ 1 \leq j, j' \leq l$ (see Algorithm 1 for the learning process).

A stored sequence retrieval process could start with any subsequence of r consecutive components and the activation of a component in the following cluster relies on the connections of r previous clusters. If the given subsequence is not the first r components of the sequence, the retrieval algorithm requires the information of the location of clusters. In Figure 2, the first three components s_1, s_2, and s_3 are shown with solid circles, and the components to be retrieved are shown with dashed circles.

The proposed retrieval procedure is sequential using a Winner-Takes-All (WTA) decision at each step. For brevity, we call this retrieval *Winner* in the rest of paper (see Algorithm 2).

3 Structures and Algorithms

The original learning and retrieval algorithms for TNN that were proposed by (Jiang et al., 2016) are reported in section 3.1. In sections 3.1.1 and 3.1.2, the newly proposed retrieval algorithms Winner-Cache and Winner-Explore are provided respectively. Feedback TNN structure along with its corresponding learning and retrieval algorithms, Feedback-Forward and Feedback-Backward, are presented in section 3.2. Finally, the error types that are used for evaluation of structures are addressed at the end of this section (section 3.3).

3.1 Learning and Retrieval Algorithms in TNN

TNN structure, which is explained in section 2.2, is summarized by Algorithm 1 and Algorithm 2 for the learning and retrieval phases respectively.

Algorithm 1: Learning in TNN

```
input : $c,\ k,\ r,\ L$ & $\mathcal{S}$
initialization
$l = 2^k$,
Generate directed graph $G$ with $n = c \times l$ nodes structured in $c$ clusters of size
  $l$.
Assign clusters indices from $1$ to $L$ cyclically (similar to Figure 2)
begin
    for $s \in \mathcal{S}$ do
        Activate the corresponding neurons to the sequence components;
        Connect each active neuron to the consecutive $r$ active neurons.
output: $G$
```

Algorithm 2: Winner Retrieval in TNN

```
input : $G$ & $[s_1 : s_r]$
initialization
Activate $r$ neurons in the first $r$ clusters using $[s_1 : s_r]$
begin
    for $i \in [r + 1 : L]$ do
        Establish the output edges from previous $r$ active neurons in the
          sequence;
        Create a candidate set of nodes with maximum score in cluster $i$.
        if len(candidate set) $==$ 1 then
            activate the only candidate node as winner and record it as $s_i$
        else if len(candidate set) $>$ 1 then
            activate one of the candidate nodes randomly as winner and record
              it as $s_i$;
output: $s_{[s_1:s_r]}$                    // Retrieved sequence given $[s_1 : s_r]$
```

For retrieval, the first r components of a previously learnt sequence, $[s_1 : s_r]$ and the learnt graph, G, are given and the complete sequence starting with $[s_1 : s_r]$ is expected.

In Algorithm 2, first each of the given r components are mapped to their related neurons in the first r clusters. Note that each component value is a number from 0 to $l - 1$. Then, the retrieval algorithm establishes the output edges from these r active neurons. The neurons in the destination cluster with highest input score will form the candidate set for the next component of the sequence. If there is just one candidate it will be added to the retrieved sequence and activated for retrieving the next component. Otherwise, the component must be chosen randomly among the candidates.

3.1.1 Winner-Cache Retrieval in TNN

In the case of Winner-Cache algorithm, the learning phase is similar, but the retrieval is more advanced. As reported in Algorithm 3, a temporary cache memory is used in the cases where random selection among winners results into an error which is detected later (see Figure 3 for an illustration).

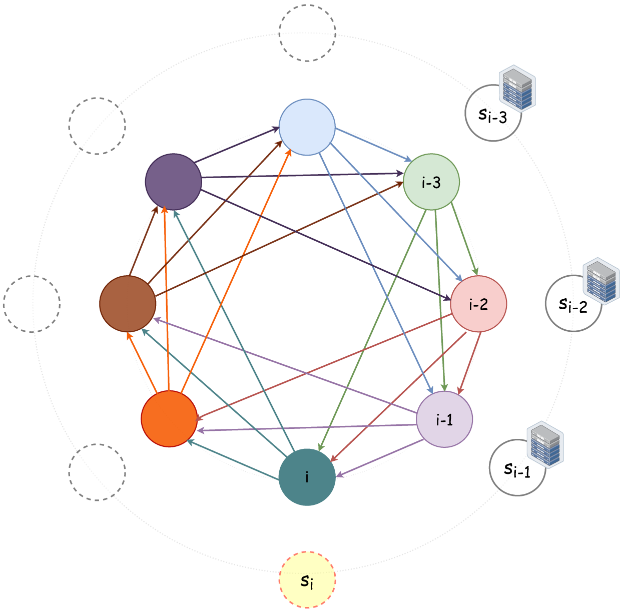


Figure 3: The mechanism of using temporary cache memory in the Winner-Cache retrieval is illustrated. The component s_i, with yellow color, represents the point in the retrieval where none of the nodes in cluster i has a score equal to $r = 3$ from last three previous activated neurons. This means that s_{i-1}, s_{i-2}, and s_{i-3} do not belong to any of previously stored sequences. Starting from cache memory in cluster $i - 3$ for component s_{i-3}, if there is an alternative candidate to be activated, we change the component, and start retrieving the sequence from that point. If in s_{i-3} the cache memory is empty, the algorithm checks for s_{i-2} and then s_{i-1}. At the end, if there is no alternative, or using the alternatives does not help, the candidate set for component s_i will be one of the winners, i.e. a node with maximum score.

The Cache-Winner algorithm proceeds as follows: whenever there is no unique candidate, the component is chosen randomly among the candidates and other candidates will be recorded temporarily (up to assignment of the next r components). If the algorithm can not find a candidate connected to all the previous r active neurons, the algorithm starts retrieval from the earliest non-empty cache memory by randomly choosing another member. For the sake of brevity, we refer to this retrieval as *Cache* in the rest of paper.

Algorithm 3: Winner-Cache Retrieval in TNN.

```
input : $G$ & $[s_1 : s_r]$
initialization
Activate $r$ neurons in the first $r$ clusters using $[s_1 : s_r]$
begin
    $i = r$
    while $i < L$ do
        $i+ = 1$
        Establish the output edges from last $r$ active neurons in the sequence;
        Create a candidate set of nodes with maximum score in cluster $i$.
        if maximum score $< r$ then
            search in the cache data of last $r$ neurons ($[i - r : i - 1]$), find the
              first non-empty cache ($j$) and select a new member randomly.
              Update the cache by removing the new member and start retrieval
              from that point ($j$) again by putting $i = j$.
        if len(candidate set) $==$ 1 then
            activate the only candidate node as winner and record it as $s_i$
        else if len(candidate set) $>$ 1 then
            activate one of the candidate nodes randomly as winner and record
              it as $s_i$;
            Put the remaining members of the candidate set into a temporary
              cache;
            keep the cached data until the next $r$ neurons are assigned.
output: $s_{[s_1:s_r]}$                // Retrieved sequence given $[s_1 : s_r]$
```

3.1.2 Winner-Explore Retrieval for TNN

At this juncture, we introduce a retrieval technique which performs exploration within the forthcoming clusters to find a more accurate solution. As reported in Algorithm 4, whenever the candidate set in a cluster is not unique, by using the previous activated neurons, we produce possible candidates in the next clusters and consequently try to eliminate the current candidates by exploring the connections to the generated candidate sets (see Figure 4 for an illustration). The maximum number of clusters that can be investigated ($r_{explore}$) is upper bounded by $r-1$. However, as will be discussed in section 4.1.1 one could limit the retrieval algorithm to explore shorter distances. For instance setting $r_{explore} < c - r$ in order to reach each cluster at most once for a specific component. Exploration involves searching for candidate sets in the following clusters and then trying to eliminate the number of candidates in the current cluster. The two techniques for this part are called Forward technique and Clique technique. In Forward technique, any candidate which is not connected to at least one node in the following clusters will be deleted from candidate set. Therefore, it is possible to find a unique candidate by reducing the size of candidate set. Clique technique is more advanced since it removes the candidates that are not in a tournament of largest possible size. We use term Clique for this technique to differentiate this technique from the learning on chain of tournaments.

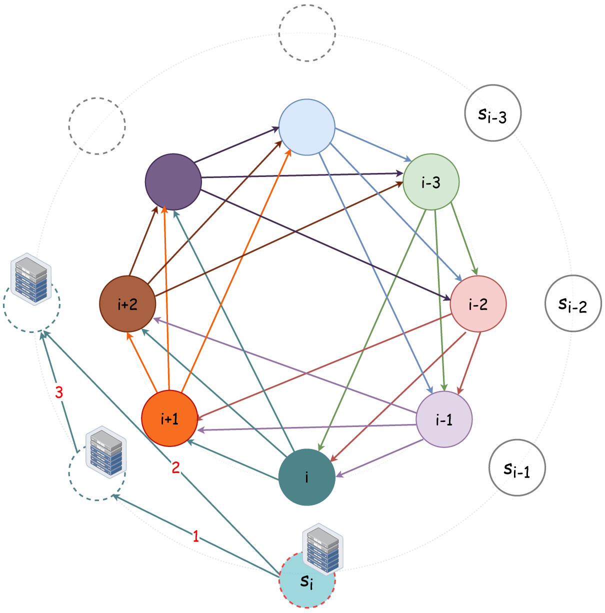


Figure 4: Using exploration technique to eliminate the number of components that are chosen randomly among the winners in Winner-Explore retrieval algorithm is illustrated. Suppose that by using the edges from $r = 3$ previous nodes equivalent to s_{i-3}, s_{i-2}, and s_{i-1} to find s_i component, more than one option is found for the candidate set in cluster i. In this case, $r_{explore} = r - 1 = 2$ previous components, i.e. s_{i-2} and s_{i-1} are used to create a candidate set in cluster $i+1$. In the Forward technique, the algorithm checks which candidates for component i are connected to at least one of the nodes in the candidate set in cluster $i+1$ (using links labeled with 1). If there is still more than one option, a candidate set in cluster $i+2$ will be constructed using s_{i-1}. Again, using Forward technique, the connections between candidates in cluster i and the candidate sets in the following $i+1$ and $i+2$ clusters are used to eliminate the options (labeled with 1 and 2). If still no unique option is available, Clique technique will be used which searches for the possible cliques of size 3 (using all the links labeled with $1, 2$, and 3). Since $r_{explore} = 2$, if there is no unique candidate in cluster i within the cliques, the process stops and the winner will be chosen randomly.

The retrieval process, as reported in Algorithm 4, searches for a candidate set in one cluster at each iteration: first by using the Forward technique, and then applying Clique technique. In the case of a non-unique option, algorithm proceeds by adding a new candidate set in the following cluster, and so on. The search for unique candidate stops whenever a unique option is found or all the clusters for exploration are taken into computation.

Algorithm 4: Winner-Explore Retrieval in TNN

```
input  : $G$ & $[s_1 : s_r], r_{explore}$
initialization
Activate $r$ neurons in the first $r$ clusters using $[s_1 : s_r]$
begin
    for $i \in [r + 1 : L]$ do
        Establish the output edges from last $r$ active neurons in the sequence;
        Create a candidate set of nodes with maximum score in cluster $i$.
        if len(candidate set) $==$ 1 then
            activate the only candidate node as winner and record it as $s_i$
        else if len(candidate set) $>$ 1 then
            for $j \in [1 : r_{explore}]$ do
                Create a candidate set in cluster $i + j$ using the $r - j$ activated
                  nodes prior to $i$;
                Construct a sub-graph of $G$ with nodes of candidate sets in
                  cluster $i$ up to cluster $i + j$;
                Update the candidate set in cluster $i$ by keeping nodes with
                  maximum output edges in sub-graph
                if len(candidate set) $==$ 1 then
                    activate the only candidate node as winner and record it as
                      $s_i$.              // Forward technique worked.
                else if len(candidate set) $>$ 1 then
                    Find all tournaments in the sub-graph including nodes from
                      candidate set in cluster $i$ with size $j + 1$;
                    Update the candidate set in cluster $i$ so that only candidates
                      in such tournaments remain;
                if len(candidate set) $==$ 1 then
                    activate the only candidate node as winner and record it as
                      $s_i$.              // Clique technique worked.
                else if len(candidate set) $==$ 0 or $j == r - 1$ then
                    Return the last non-empty candidate set as the final
                      candidate set for cluster $i$;
output: $s_{[s_1:s_r]}$              // Retrieved sequence given $[s_1 : s_r]$
```

13

3.2 Feedback TNN Structure

In this structure, the learning phase sets tournaments with forward and backward connections. Each node in a tournament of size $r + 1$, has r_{fwd} links to the forthcoming clusters and receives r_{fbk} links from the forthcoming $[r_{fwd} + 1 : r]$ active neurons, where $0 \geq r_{fbk} \geq r_{fwd}$ and $r = r_{fwd} + r_{fbk}$ (see Figure 5).The original TNN can be seen as a Feedback TNN with zero feedback links ($r_{fbk} = 0$).

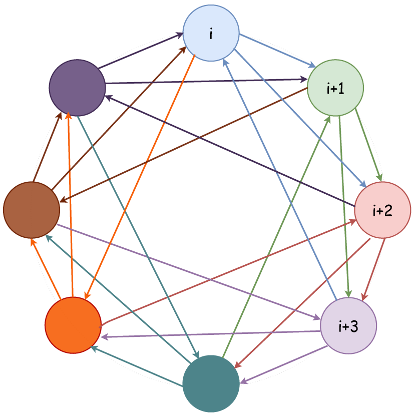


Figure 5: In the chain of tournament structure with feedback connections, the first r_{fwd} connections of each tournament are clockwise and the next r_{fbk} connections are counterclockwise. In this illustration, $r_{fwd} = 2$ and $r_{fbk} = 1$.

For storing sequence $s \in \mathcal{S}$, where $s = s_1 s_2 \cdots s_L$, the clockwise connections in the network will be as follows:

$$n_{ij} \rightarrow n_{i'j'}, \text{ if: } \begin{cases} f_i(s_{i+(\pi-1)c}) = n_{ij} \\ f_{i'}(s_{i'+(\pi-1)c}) = n_{i'j'} \end{cases} \text{and, } 1 \leq \delta_i(i') \leq r_{fwd}, \tag{3}$$

and for counterclockwise connections:

$$n_{ij} \leftarrow n_{i'j'}, \text{ if: } \begin{cases} f_i(s_{i+(\pi-1)c}) = n_{ij} \\ f_{i'}(s_{i'+(\pi-1)c}) = n_{i'j'} \end{cases} \text{and, } r_{fwd} \leq \delta_i(i') \leq r \tag{4}$$

where $1 \leq \pi \leq \lfloor \frac{L}{c} \rfloor$. In general, for $s \in \mathcal{S}$, if the above conditions are satisfied for a given π such that $n_{ij} \rightarrow n_{i'j'}$, we set $N_{s,\pi}(n_{ij}, n_{i'j'}) = 1$. Similarly we set $N_{s,\pi}(n_{i'j'}, n_{ij}) = 1$, if $n_{ij} \leftarrow n_{i'j'}$; otherwise we set $N_{s,\pi}(n_{ij}, n_{i'j'}) = 0$, and $N_{s,\pi}(n_{i'j'}, n_{ij}) = 0$.

At the end of learning or storing process, the network has the following connections:

$$\mathcal{W} = \{\omega_{(ij)(i'j')} | \text{ if } \exists\, s \in \mathcal{S}, \text{ and } \exists\, \pi \in [1 : \lfloor \frac{L}{c} \rfloor] \text{ s.t. } N_{s,\pi}(n_{ij}, n_{i'j'}) = 1\} \tag{5}$$

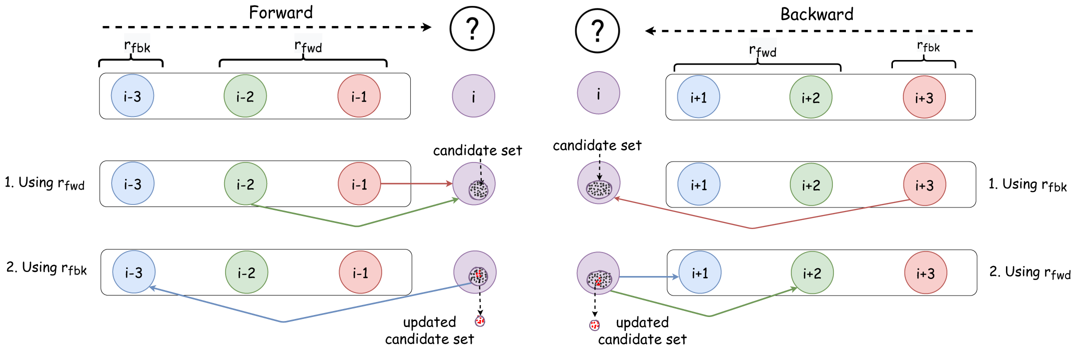
where $\omega_{(ij)(i'j')}$ is a directed edge from n_{ij} to $n_{i'j'}$ and $1 \leq i, i' \leq c$, $1 \leq j, j' \leq l$ (see Algorithm 5 for the learning process). In Figure 5, activated neurons in cluster i are connected to the activated neurons in clusters $i + 1$ and $i + 2$ clockwise, whereas activated neurons in cluster $i + 3$ are connected to the activated neurons in cluster i counterclockwise.

Algorithm 5: Learning in Feedback TNN

```
input : $c, k, r, r_{fwd}, L$ & $\mathcal{S}$
initialization
$l = 2^k,\ r_{fbk} = r - r_{fwd}$
Generate directed graph $G$ with $n = c \times l$ nodes structured in $c$ clusters of size
  $l$.
Assign clusters indices from $1$ to $L$ cyclically (see Figure 5 for labeling)
begin
    for $s \in \mathcal{S}$ do
        Activate the corresponding neurons to the sequence;
        Connect each active neuron (say in cluster $i$) to the active neurons in
          the next $r_{fwd}$ clusters ($[i + 1 : i + r_{fwd}]$);
        Connect each active neuron to the previous $r_{fbk}$ active neurons in
          clusters $[i - r : i - r_{fwd} - 1]$;
output: $G$
```

3.2.1 Retrieval in Feedback TNN

Here we introduce two retrieval algorithms, Feedback-Forward (Algorithm 6) and Feedback-Backward (Algorithm 7), which can retrieve a complete sequence from any given segment. To do so, we need a pre-matching process to find the clusters on which the given sequence segment was stored (see Figure 6 for an illustration of Feedback-Forward and Feedback-Backward processes).



(a) For Forward retrieval in Feedback TNN, first a candidate set in cluster i is created using the connections from active neuron in clusters $i-1$ and $i-2$ (since $r_{fwd}=2$). If there is a unique winner candidate, the algorithm stops, otherwise a sub-graph is constructed with the candidate set and the active neuron in cluster $i-3$ (since $r_{fbk}=1$). The candidate set will be updated by keeping nodes with maximum score.

(b) For Backward retrieval in Feedback TNN, first a candidate set in cluster i is created using the connections from active neuron at cluster $i+3$ (since $r_{fbk}=1$). If there is a unique winner candidate, the algorithm stops, otherwise a sub-graph is constructed with the candidate set and the active neurons in clusters $i+1$ and $i+2$ (since $r_{fwd}=2$). The candidate set will be updated by keeping nodes with maximum score.

Figure 6: Consider the structure in Figure 5 where $r_{fwd}=2$ and $r_{fbk}=1$. Given a segment of $r=3$ components, Forward and Backward retrieval processes are illustrated respectively in (a) and (b).

Feedback-Forward algorithm (hereafter Forward) retrieves the sequence given the first r components of it. This retrieval is performed in two phases: first, by using the r_{fwd} connections, and then if the winning candidate is not unique, the r_{fbk} connections are used to eliminate the number of candidates, as reported in Algorithm 6.

Feedback-Backward algorithm (hereafter Backward), retrieves the sequence given the last r components of a sequence. As reported in Algorithm 7, the algorithm first uses the r_{fbk} input edges to make an initial candidate set, and then the output edges from the candidate set is used to eliminate the number of candidates.

Algorithm 6: Feedback-Forward Retrieval in Feedback TNN

```
input  : $G$ & $[s_1 : s_r]$
initialization
Activate $r$ neurons in the first $r$ clusters using $[s_1 : s_r]$
Assign clusters indices from $1$ to $L$ cyclically
begin
    for $i \in [r + 1 : L]$ do
        Establish the output edges from previous $r_{fwd}$ active neurons in the
          sequence;
        Create a candidate set of nodes with maximum score in cluster $i$.
        if len(candidate set) $==$ 1 then
            activate the only candidate node as winner and record it as $s_i$
        else if len(candidate set) $>$ 1 then
            A sub-graph of $G$ with nodes from candidate set in cluster $i$, and
              previous $r_{fbk}$ active neurons in clusters $[i - r : i - r_{fwd}]$ is
              constructed;
            The new candidate set for cluster $i$ is updated by keeping the nodes
              which have maximum output edges in the sub-graph;
            Select one node from the updated candidate set as winner and
              record it as $s_i$;
output: $s_{[s_1:s_r]}$                    // Retrieved sequence given $[s_1 : s_r]$
```

Algorithm 7: Feedback-Backward retrieval in Feedback TNN.

```
input : $G$ & $[s_{L-r+1} : s_L]$
initialization
Activate $r$ neurons in the related $r$ clusters using $[s_{L-r+1} : s_L]$
Assign clusters indices from $1$ to $L$ cyclically
begin
    for $i \in [L - r : 1; -1]$ do
        Establish the output edges from $r_{fdk}$ active neurons in clusters
          $[i + r_{fwd} : i + r]$;
        Create a candidate set of nodes with maximum score in cluster $i$.
        if len(candidate set) $==$ 1 then
            activate the only candidate node as winner and record it as $s_i$
        else if len(candidate set) $>$ 1 then
            A sub-graph of $G$ with nodes from candidate set in cluster $i$, and the
              next $r_{fwd}$ active neurons is constructed;
            The new candidate set for cluster $i$ is updated by keeping the nodes
              with maximum score (maximum output edges) in the sub-graph;
            Select one node from the updated candidate set as winner and
              record it as $s_i$;
output: $s_{[s_{L-r+1}:s_L]}$          // Retrieved sequence given $[s_{L-r+1} : s_L]$
```

Note that Winner (Algorithm 2) can be seen as a special case of Forward (Algorithm 6) when $r_{fwd} = r$ and $r_{fbk} = 0$. In Figure 6a, only the first step that uses r_{fwd} is applicable. On the other hand, in the case of the original TNN, the Backward algorithm starts with a candidate set of size l and makes a sub-graph with the given $r_{fwd} = r$ components, since $r_{fbk} = 0$ and there is no input connection. In Figure 6b, only the second step that uses r_{fwd} is applicable.

3.3 Error Types

Based on the argument of Jiang et al. (2016), two different error types could be distinguished; an error type that is due to prior retrieval errors in simulation, and an error type that is structural and which is caused by an excessive network density. The structural error type could happen even if all the previous r components are given correctly.

Component Error Rate (CER) and Sequence Error Rate (SER) address the simulation error; CER is defined as the ratio of the number of incorrect components over the number of total retrieved components, whereas SER is defined as the number of sequences that are failed to be retrieved correctly over the total number of sequences.

Structural Component Error Rate (S-CER) and Structural Sequence Error Rate (S-SER) address the structural error. According to Jiang et al. (2016), the S-CER can be estimated as the error rate at a single retrieval step when the provided previous r components are correct.

$$P_{S-CER} = 1 - (1 - d^r)^{l-1} \tag{6}$$

where d is the network density which is the ratio of number of established connections during the storage process over all possible connections that the network structure allows. The density is calculated (in Jiang et al., 2016, equation 7) as:

$$d = 1 - \left(1 - \frac{1}{l^2}\right)^{\frac{|S|L}{c}} \tag{7}$$

At the sequence level, S-SER is estimated (in Jiang et al., 2016, equation 9) as:

$$P_{S-SER} = 1 - (1 - d^r)^{(l-1)(L-r)} \tag{8}$$

Please note that the density in the Feedback TNN structure is the same as the density of the original TNN structure (equation 7). This is due to the fact that the density is calculated based on the probability of having a connection between two nodes, and in the case of Feedback TNN just the directions of some connections are changed while their number remains the same. Moreover, based on the definition of structural errors, equations 6 and 8 are valid for Cache, Explore and Feedback TNN retrievals.

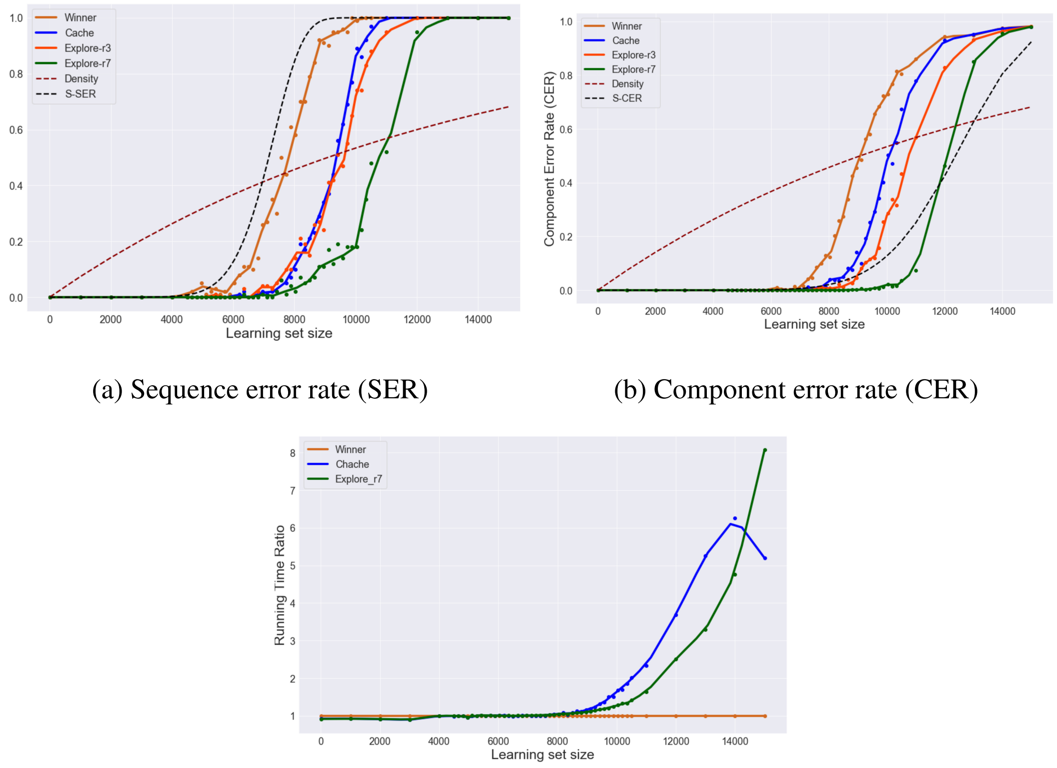
4 Simulation Results

In this section, the simulation results for different algorithms are presented in order to show the robustness of storage and to compare different structures. Learning processes for TNN and Feedback TNN structures (Algorithm 1 and Algorithm 5, respectively) are considered when $c = 20$, $k = 8$, $l = 2^8 = 256$, $r = 12$, $r_{fwd} = 6$, $r_{fbk} = r - r_{fwd} = 6$, and $L = 100$. Regarding the retrieval, four scenarios; *Winner* (Algorithm 2), *Cache* (Algorithm 3), *Explore* (Algorithm 4), *Forward* (Algorithm 6), and *Backward* (Algorithm 7) are simulated and compared.

The sequences in the learning set are different in at least one of the first r components. For instance, a learning set of size 1000 is a set of 1000 sequences that all are different in at least one component in the 12 first components. To see if the memorized sequences can be retrieved, 100 of the learnt sequences are randomly chosen from each learning set. To reduce randomness effect, we fixed the 100 choices of sequences in the learning set of each size (varies between 10 to 15000), in simulations for all the retrieval algorithms.

4.1 TNN Retrieval Results

Figure 7 depicts the error rate for a range of learning set sizes, for different retrieval algorithms, namely, Winner (Algorithm 2), Cache (Algorithm 3) , Explore with $r_{explore} = 3$, and $r_{explore} = 7$ (Algorithm 4). To illustrate the power of the algorithms with respect to the structure of the network, the calculated density and structured error are also plotted. It is clear from the results that retrieval with the exploration when $r_{explore} = 7$ is far better than the rest of scenarios. For instance, when the learning set is composed of 10000 sequences, each of size 100, the SER (Figure 7a) for the Winner is one, which means that no sequence can be retrieved correctly with the original algorithm. While this value is about 0.7 for the algorithm with cache memory and about 0.6 when the exploration technique is used with $r_{explore} = 3$, and the SER for exploration with $r_{explore} = 7$ is less than 0.2. This superiority of exploration algorithm can easily be tracked in the CER results (Figure 7b). For instance, for the same learning set, the CER for Winner is 0.75, for Cache it is 0.4, for Explore with $r_{explore} = 3$ it equals to 0.3, and for Explore with $r_{explore} = 7$ it is near zero.



(a) Sequence error rate (SER)

(b) Component error rate (CER)

(c) Running time ratio for Explore-$r7$ and Cache over Winner.

Figure 7: Comparison between retrieval algorithms on the TNN structure; Winner (Algorithm 2), Cache (Algorithm 3) , Explore with $r_{explore} = 3$, and $r_{explore} = 7$ (Algorithm 4). The running time ratios of Explore (with $r_{explore} = 7$) and Cache over Winner are reported in 7c.

In Figure 7a, the simulated error value for all retrieval methods are less than S-SER which is obtained from equation 8. This can be explained by the fact that the S-SER error estimation is based on the probability of having at least two nodes in a cluster that all the previous r components are connected to. In this case, for the simplest version of retrieval algorithms, Winner, one candidate will be chosen randomly. In other words, S-SER is an upper bound for SER and in the case that all the choices are unique (S-SER $= 0$), there will be no error (SER $= 0$). Although there is no guarantee that the randomly chosen candidate is the desired one, the SER value is slightly less than the S-SER. Obviously, the more sophisticated retrieval algorithms, Cache and Explore, reduce the random selections and therefore, the number of errors. The structure error is a function of network density and as can be seen in Figure 7, higher density leads to higher structure error.

For S-CER (Figure 7b), the argument is different and the simulated error values in retrieval process are higher than S-CER. To calculate S-CER, the assumption is that the previous retrieved components are correct and S-CER estimates the probability of having at least two nodes that are fully connected to the previous r components. However, in the simulation, the values of some of r previous components are faulty and as a result the decision is not based on correct components. Therefore, in a sequence retrieval, errors at each component could be propagated to the rest of retrieval and simulated error CER will be higher than S-CER which assumes the r components are correct.

The reported results in Figure 7 suggest Explore retrieval with higher number of steps. Cache algorithm is also promising, but for large learning sets it has a low speed. When the network density increases, Cache retrieval process creates larger candidate sets for each component and therefore larger cache memory, and the algorithm might go through all the options to find the correct component. Explore, on the other hand, must explore longer distances that is the source of complexity in Explore. Figure 7c compares the simulation running time between Explore-$r7$ and Cache with Winner for different learning set sizes. The running time up to a learning set size of 8000 for all the three algorithms is the same, while Explore-$r7$ and Cache perform far better than Winner; compare the low performance of Winner ($SER = 0.52$) with the performance of Cache ($SER = 0.08$) and Explore-$r7$ ($SER = 0.02$). As another example, for learning set size 10000, $SER = 1$ for Winner; while Explore-$r7$ has $SER = 0.18$ and running time ratio 1.2, and Cache has $SER = 0.86$ and running time ratio 1.7.

This shows that for reasonable error values (say less than 0.1), the running time ratio is at the same level of Winner in both cases. Interestingly, the running time for Cache reaches a peak for a learning set of size 14000 and thereafter starts to decline for larger learning sets as shown in Figure 7c. This can be explained by the excessive density so that the probability of having full score candidate at each step increases and therefore the algorithm can not detect an error which reduces the processing time for checking the Cache memory.

In Figure 7b, only Explore algorithm with $r_{explore} = 7$ that investigates further clusters shows lower error than S-CER until the density about 0.6 and learning set of size 12000. We will have a closer look at the simulation results for the Explore algorithm below.

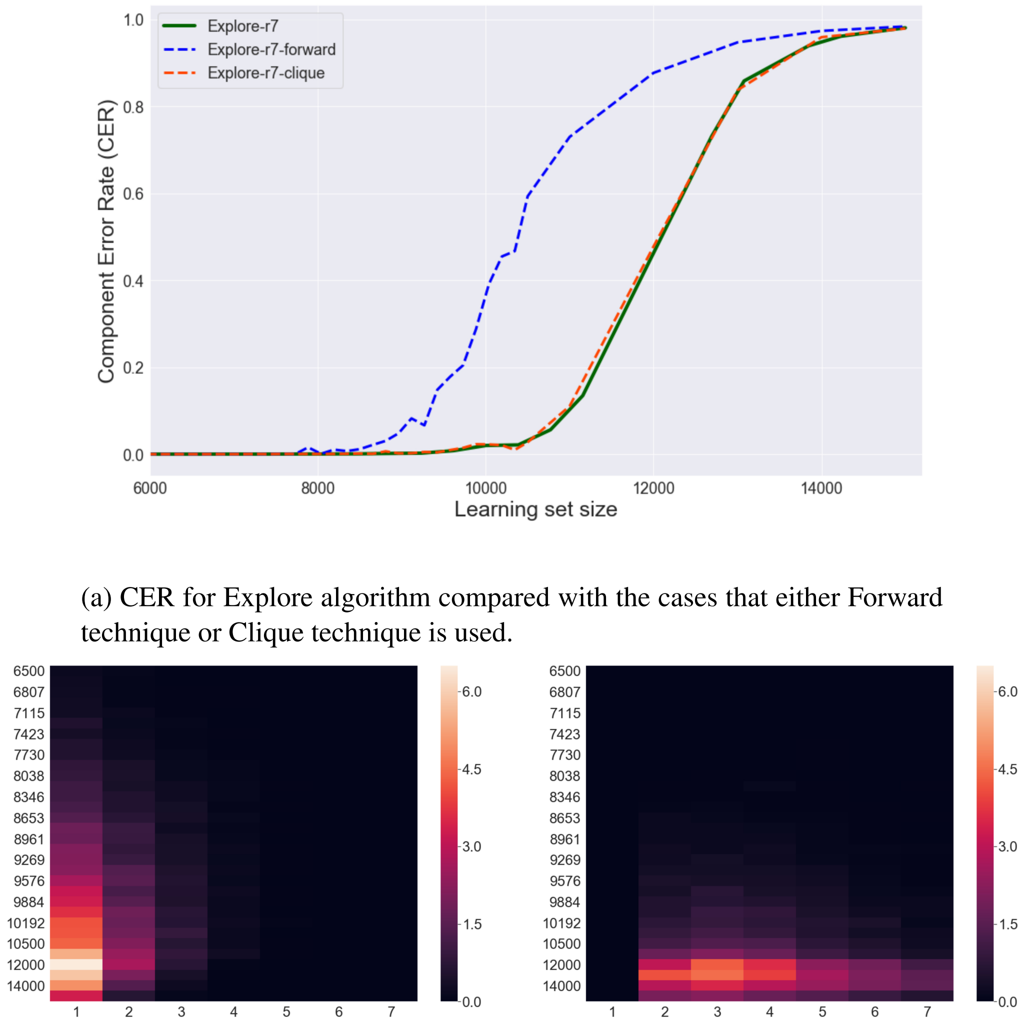
4.1.1 More Investigation on Explore Retrieval Algorithm

In Explore retrieval, by starting from distance one, the algorithm uses Forward and Clique techniques consecutively and increases the exploration distance until a unique candidate is found or $r_{explore}$ limit is met. Clique technique is more powerful but it is more computationally expensive than Forward technique. Figure 8a shows that by using the Clique technique alone (red dashed line) the exploration performance does not change, whilst Forward technique alone (blue dashed line) is far less effective than the achieved results by exploration algorithm. This is an expected result since Clique technique is endowed with Forward technique.

Figure 8b and 8c show the number of components that Forward and Clique techniques successfully retrieved (unique winner), respectively in the course of retrieving each sequence. The columns show the exploration distance and the rows show the size of learning sets. It is noteworthy that the first column in Figures 8c is all zero since for a distance one, a forward connection and a tournament of size 2 are the same, and the Forward technique is prior to the Clique technique in Algorithm 6.

As reported in Figure 7b, the Winner handles the retrieval when the learning set sizes are up to 7000. Until this point, no exploration is demanded. But with larger sizes of learning set and whenever it comes to the exploration phase, most of the cases can be retrieved with exploration of distance one. This, however, does not mean that the best choice, in terms of time/accuracy trade off, is $r_{explore} = 1$. When the size of learning sets gets higher, the Clique technique gets more involved. Because the higher sizes of candidate sets in under exploration clusters increases the searching domain, which results Forward technique to be failed in retrieval and Clique technique starts to retrieve. Let us consider for instance the learning set sizes around $12000 - 13000$ which is the highest number of successful retrievals per sequence using Explore retrieval (Figure 7a). For these sizes the CER error is high, for example it is about 0.46 for learning set of size 12000 and equals 0.85 when the learning set size is 13000 and therefore the overall retrieval is not successful. Interestingly, the S-CER also beats CER at around 12000 (Figure 7b) which shows that high density can not be managed with exploration technique as well.

For learning sets of size 11000, the CER for Explore-$r7$ is 0.074 (Figure 7b) while without exploration technique the CER value equals one for learning set sizes larger than 10000. Figure 8b and 8c show decrease in the successful cases at exploration with higher distances, say 6 or 7 which suggests that extra exploration is not worth the computation. We found $r_{explore} = 7$ as a suitable choice for this setting of parameters.



(a) CER for Explore algorithm compared with the cases that either Forward technique or Clique technique is used.

(b) Number of unique winner components which are found at $[1:7]$ exploration distances using Forward technique.

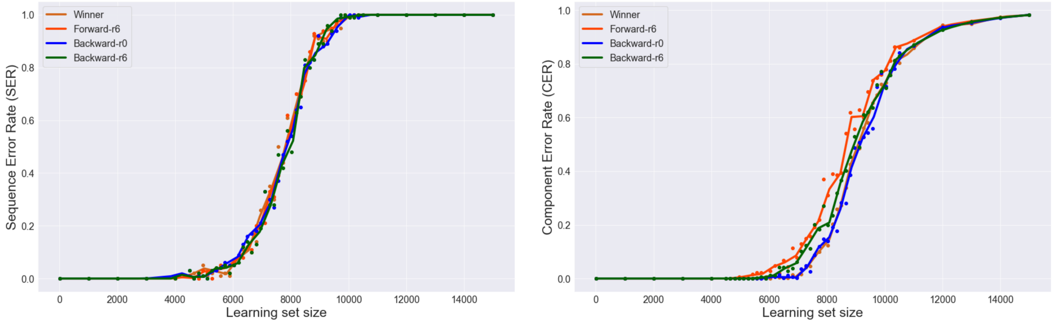
(c) Number of unique winner components which are found at $[1:7]$ exploration distances using Clique technique (with tournament sizes $[1:7]+1$)

Figure 8: Analysis of Explore retrieval; Forward technique vs. Clique technique and the required exploration distance for finding a unique component. Results of learning set sizes between 6000 and 15000 are depicted.

4.2 Feedback TNN Retrieval Results

Figure 9 shows the retrieval error of Feedback TNN learning when $r = 12$ & $r_{fwd} = 6$ (Forward-$r6$ and Backward-$r6$) together with the retrieval error of original learning method (TNN) with Winner and Backward-$r0$ retrievals when $r = 12$. We start the Winner and Forward-$r6$ retrievals when the first $r = 12$ components are given, and Backward-$r6$ and Backward-$r0$ when the last $r = 12$ components are given.

Figure 9a confirms that the sequence retrieval results in Feedback TNN can be as accurate as the original TNN memories. It is almost the same for CER (Figure 9b), however the results for the original TNNs are slightly better. We can explain this as a result of errors in recent previous $r_{fwd} = 6$ components. Consider the case that the algorithm finds a unique candidate for the current component based on last $r_{fwd} = 6$ components, without considering the other $r_{fbk} = 6$ links, and selects it as the only winner, while it can be incorrect candidate due to some errors in previous steps. However, if the algorithm uses all the r_{fwd} and r_{fbk} links the candidate set might composed of more components, which are not necessarily of full score. In this case, the final candidate will be chosen randomly, and therefore there is a chance of correct component selection. The above argument could similarly explain why CER for Backward-$r0$ are slightly better than Backward-$r6$. Note that the errors in Feedback TNN retrievals might cause more random choices in retrieval of the rest of components (see section 4.3 for an analysis of randomly chosen components). Indeed, such errors do not increase SER but CER could be affected as seen in Figure 9b.



(a) Sequence error rate

(b) Component error rate

Figure 9: Comparison between the original TNN learning method and the learning in Feedback TNN using Winner, Forward and Backward retrievals.

In summary, in Feedback TNN the retrieval is faster than TNN, the SER performance is the same for both, but TNN could be slightly better in CER performance.

4.3 Randomness in Simulated Retrievals; an Overall Look

Figure 10 provides a general overview on the number of cases in average that retrieval algorithms select the final component randomly from the candidate set. The success in policy of reducing the number of cases with random decision in Cache and Explore retrievals to achieve better retrieval performance is clearly shown in the last three columns related to these retrievals. For instance, when the learning set size equals 11000, nearly 50 components out of $L - r = 88$ are chosen randomly for Winner, as the original retrieval algorithm, but it is about 20 for Cache, 15 for Explore with $r_{explore} = 3$, and almost zero for Explore with $r_{explore} = 7$. The number of random choices for Feedback TNN structure, both Forward and Backward, is slightly higher than Winner and Backward-$r0$. The argument is that the errors that appear due to the wrong unique retrieval, produce more error afterwards in the sequence, and therefore more random winner retrieval cases in total. We also can observe a slightly higher number of random winner selection in the Backward-$r0$. This could be related to the learning set generation in our simulations. The sequences in a learning set, are forced to be different in at least one of the first r components. Therefore, the Winner can start the retrieval with the unique sequence, while in the Backward-$r0$ more than one sequence can match with the given last r components.

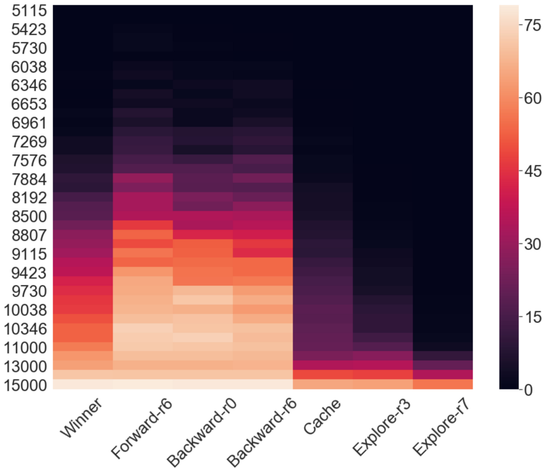


Figure 10: A comparison between number of random selection of winner candidate in different scenarios.

5 Discussion and Concluding Remarks

In this study, two-fold contributions within the field of TNN structures were presented; first, we proposed a more general learning and retrieval structure called Feedback TNN, and second, we devised two more accurate retrieval algorithms in comparison with the Winner algorithm.

In Feedback TNN, each segment of sequence of length $r+1$ is mapped into a tournament in $r+1$ consecutive clusters where each neuron has r_{fbk} input edges and $r_{fwd} = r - r_{fbk}$ output edges. The proposed retrieval for the Feedback TNN operates in two phases, in a faster manner than TNN retrieval, and generates the same sequence error rate while producing a slightly weaker component error rate.

The original TNN can be considered as a special case of Feedback TNN with zero feedback connections. Using feedback connections, we obtained results of sequence retrieval as precise as the original structure, with the possibility of faster retrieval. One might also divide the r forward connections into two parts, say r_1 and r_2, and try to retrieve the component using the most recent r_1 active neurons, and if it is not possible to uniquely retrieve, use the rest of r_2 neurons. More generally, one can try to retrieve by starting from the last active neuron and reduce the size of the candidate set (losers-kicks-out), and adding more active neurons to the retrieval process, until either one winner candidate remains or all the r active neurons are used.

By introducing Backward retrieval in this paper, we showed that it is possible to get a part of a sequence, no matter its location, and retrieve the rest. In this case, the retrieval algorithm must be able to first locate a tournament matching the given sub-sequence, and later retrieve the whole sequence from both directions. Backward retrieval is compatible with both TNN and Feedback TNN structures, but Feedback TNN with non-zero feedback links is preferable since the Backward retrieval algorithm can start with a smaller size candidate set.

In order to improve the retrieval accuracy for a given network, we suggested two algorithms with the overall strategy to limit the number of random selections during retrieval. The Cache retrieval (Algorithm 3) uses a temporary cache memory for the last r components to record the candidate set of winners whenever the chosen winner is not unique. These cached alternatives are used whenever the algorithm detects an error by observing no candidate having a full score. The reported results in section 4 confirm the usefulness of this method. The more advanced, and successful, retrieval algorithm (Algorithm 4) explores the forthcoming clusters to find a unique candidate in the current cluster. This algorithm somehow investigates the consequence of choosing each candidate by checking its connections to the possible future components and decides more judiciously. This algorithm produces the best results.

Explore-Winner is a more reliable retrieval method than Cache-Winner since it limits the number of random choices using the data in the forthcoming clusters, while Cache-Winner tries to correct the errors by testing other possibilities. Cache-Winner might be computationally expensive in higher densities where candidate sets of winners are larger and therefore, larger sets are cached. Finding an optimal $r_{explore}$, for exploration distance limit, as shown in section 4.1.1, is a trade-off between time and accuracy. Although not reported in the simulations, both Cache-Winner and Explore-Winner can be used in Feedback TNN and for Backward retrieval.

Similar to the double-layer structure proposed by Jiang et al. (2016), it is possible to consider a hierarchical structure by adding an extra connectivity level. Moreover, similar to the technique used in (Mofrad et al., 2016) a precoding could dramatically increase the storage and retrieval capacity by forcing patterns to be well separated and therefore reducing the common tournaments in different patterns.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback who helped to improve the quality of the manuscript. The source code of the simulations is made publicly available online under this link: `https://github.com/Asieh-A-Mofrad/Tournament-Based-Sequence-Storage`.

References

Aboudib, A., Gripon, V., & Coppin, G. (2016). A neural network model for solving the feature correspondence problem. In *International Conference on Artificial Neural Networks* (pp. 439–446).: Springer.

Aboudib, A., Gripon, V., & Jiang, X. (2014). A study of retrieval algorithms of sparse messages in networks of neural cliques. In *COGNITIVE 2014: the 6th International Conference on Advanced Cognitive Technologies and Applications* (pp. 140–146).

Aliabadi, B. K., Berrou, C., Gripon, V., & Jiang, X. (2014). Storing sparse messages in networks of neural cliques. *IEEE Transactions on neural networks and learning systems*, 25(5), 980–989.

Berrou, C., Dufor, O., Gripon, V., & Jiang, X. (2014). Information, noise, coding, modulation: What about the brain? In *Turbo Codes and Iterative Information Processing (ISTC), 2014 8th International Symposium on* (pp. 167–172).: IEEE.

Berrou, C. & Gripon, V. (2010). Coded hopfield networks. In *Turbo Codes and Iterative Information Processing (ISTC), 2010 6th International Symposium on* (pp. 1–5).: IEEE.

Berrou, C. & Kim-Dufor, D.-H. (2018). A connectionist model of reading with error correction properties. In *Human Language Technology. Challenges for Computer Science and Linguistics: 7th Language and Technology Conference, LTC 2015, Poznań, Poland, November 27-29, 2015, Revised Selected Papers*, volume 10930 (pp. 304).: Springer.

Boguslawski, B., Gripon, V., Seguin, F., & Heitzmann, F. (2014). Huffman coding for storing non-uniformly distributed messages in networks of neural cliques. In *AAAI 2014: the 28th Conference on Artificial Intelligence*, volume 1 (pp. 262–268).

Brea, J., Senn, W., & Pfister, J.-P. (2011). Sequence learning with hidden units in spiking neural networks. In *Advances in neural information processing systems* (pp. 1422–1430).

Danilo, R., Jarollahi, H., Gripon, V., Coussy, P., Conde-Canencia, L., & Gross, W. J. (2015). Algorithm and implementation of an associative memory for oriented edge detection using improved clustered neural networks. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 2501–2504).: IEEE.

Gripon, V. & Berrou, C. (2011). Sparse neural networks with large learning diversity. *IEEE Transactions on Neural Networks*, 22(7), 1087–1096.

Gripon, V. & Berrou, C. (2012). Nearly-optimal associative memories based on distributed constant weight codes. In *Information Theory and Applications Workshop (ITA), 2012* (pp. 269–273).: IEEE.

Gripon, V., Heusel, J., Löwe, M., & Vermet, F. (2016). A comparative study of sparse associative memories. *Journal of Statistical Physics*, 164(1), 105–129.

Hacene, G. B., Gripon, V., Farrugia, N., Arzel, M., & Jezequel, M. (2017). Finding all matches in a database using binary neural networks. *COGNTIVE 2017*, (pp.6̃7).

Hacene, G. B., Gripon, V., Farrugia, N., Arzel, M., & Jezequel, M. (2019). Budget restricted incremental learning with pre-trained convolutional neural networks and binary associative memories. *Journal of Signal Processing Systems*, 91(9), 1063–1073.

Hawkins, J. & Ahmad, S. (2016). Why neurons have thousands of synapses, a theory of sequence memory in neocortex. *Frontiers in neural circuits*, 10, 23.

Hawkins, J. & Blakeslee, S. (2007). *On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines*. Macmillan.

Hawkins, J., George, D., & Niemasik, J. (2009). Sequence memory for prediction, inference and behaviour. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1521), 1203–1209.

Hoffmann, H. (2019). Sparse associative memory. *Neural computation*, 31(5), 998–1014.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8), 2554–2558.

Hopfield, J. J. (2008). Searching for memories, sudoku, implicit check bits, and the iterative use of not-always-correct rapid neural computation. *Neural Computation*, 20(5), 1119–1164.

Jarollahi, H., Gripon, V., Onizawa, N., & Gross, W. J. (2015). Algorithm and architecture for a low-power content-addressable memory based on sparse clustered networks. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 23(4), 642–653.

Jarollahi, H., Onizawa, N., Gripon, V., & Gross, W. J. (2014). Algorithm and architecture of fully-parallel associative memories based on sparse clustered networks. *Journal of Signal Processing Systems*, 76(3), 235–247.

Jiang, X. (2014). *Storing sequences in binary neural networks with high efficiency*. PhD thesis, Télécom Bretagne, Université de Bretagne Occidentale.

Jiang, X., Gripon, V., Berrou, C., & Rabbat, M. (2016). Storing sequences in binary tournament-based neural networks. *IEEE transactions on neural networks and learning systems*, 27(5), 913–925.

Jiang, X., Marques, M. R. S., Kirsch, P.-J., & Berrou, C. (2015). Improved retrieval for challenging scenarios in clique-based neural networks. In *International Work-Conference on Artificial Neural Networks* (pp. 400–414).: Springer.

Kim, D.-H., Park, J., & Kahng, B. (2017). Enhanced storage capacity with errors in scale-free hopfield neural networks: An analytical study. *PloS one*, 12(10), e0184683.

Krotov, D. & Hopfield, J. J. (2016). Dense associative memory for pattern recognition. In *Advances in neural information processing systems* (pp. 1172–1180).

Larras, B., Chollet, P., Lahuec, C., Seguin, F., & Arzel, M. (2018). A fully flexible circuit implementation of clique-based neural networks in 65-nm cmos. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 66(5), 1704–1715.

Larras, B. & Frappé, A. (2020). On the distribution of clique-based neural networks for edge ai. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*.

Marques, M. R. S., Hacene, G. B., Lassance, C. E. R. K., & Horrein, P.-H. (2017). Large-scale memory of sequences using binary sparse neural networks on gpu. In *2017 International Conference on High Performance Computing & Simulation (HPCS)* (pp. 553–559).: IEEE.

Maurer, A., Hersch, M., & Billard, A. G. (2005). Extended hopfield network for sequence learning: Application to gesture recognition. In *International Conference on Artificial Neural Networks* (pp. 493–498).: Springer.

Mofrad, A. A., Ferdosi, Z., Parker, M. G., & Tadayon, M. H. (2015). Neural network associative memories with local coding. In *Information Theory (CWIT), 2015 IEEE 14th Canadian Workshop on* (pp. 178–181).: IEEE.

Mofrad, A. A. & Parker, M. G. (2017). Nested-clique network model of neural associative memory. *Neural Computation*.

Mofrad, A. A., Parker, M. G., Ferdosi, Z., & Tadayon, M. H. (2016). Clique based neural associative memories with local coding and pre-coding. *Neural Computation*, 28, 1–21.

Olshausen, B. A. & Field, D. J. (2004). Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4), 481–487.

Rinkus, G. J. (2010). A cortical sparse distributed coding model linking mini-and macrocolumn-scale functionality. *Frontiers in neuroanatomy*, 4, 17.

Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature*.

Yao, Z., Gripon, V., & Rabbat, M. (2014). A gpu-based associative memory using sparse neural networks. In *2014 International Conference on High Performance Computing & Simulation (HPCS)* (pp. 688–692).: IEEE.

PAPER E

ENHANCED EQUIVALENCE PROJECTIVE SIMULATION: A FRAMEWORK FOR MODELING FORMATION OF STIMULUS EQUIVALENCE CLASSES

Mofrad, Asieh A., Yazidi, Anis, Mofrad, Samaneh A., Hammer, Hugo L. and Arntzen, Erik. 3 March 2021. Enhanced equivalence projective simulation: A framework for modeling formation of stimulus equivalence classes. *Neural Computation*. MIT Press.

Enhanced Equivalence Projective Simulation: A Framework for Modeling Formation of Stimulus Equivalence Classes

Asieh Abolpou Mofrad
asieh.abolpour-mofrad@oslomet.no
Anis Yazidi
Anis.Yazidi@oslomet.no
Department of Computer Science, Oslo Metropolitan University,
0130 Oslo, Norway

Samaneh Abolpour Mofrad
Samaneh.Abolpour.Mofrad@hvl.no
Department of Computer Science, Electrical Engineering, and Mathematical Sciences,
Western Norway University of Applied Sciences, 5063 Bergen, Norway, and Mohn
Medical Imaging and Visualization Center, Department of Radiology, Haukeland
University Hospital, 5021 Bergen, Norway

Hugo L. Hammer
Hugo.Hammer@oslomet.no
Department of Computer Science, Oslo Metropolitan University, 0130 Oslo,
Norway, and Simula Metropolitan Center, 1325 Oslo, Norway

Erik Arntzen
erik.arntzen@equivalence.net
Department of Behavioral Science, Oslo Metropolitan University,
0130 Oslo, Norway

Formation of stimulus equivalence classes has been recently modeled through equivalence projective simulation (EPS), a modified version of a projective simulation (PS) learning agent. PS is endowed with an episodic memory that resembles the internal representation in the brain and the concept of cognitive maps. PS flexibility and interpretability enable the EPS model and, consequently the model we explore in this letter, to simulate a broad range of behaviors in matching-to-sample experiments. The episodic memory, the basis for agent decision making, is formed during the training phase. Derived relations in the EPS model that are not trained directly but can be established via the network's connections are computed on demand during the test phase trials by likelihood reasoning. In this letter, we investigate the formation of derived relations in the EPS model using network enhancement (NE), an iterative diffusion process, that yields an offline approach to the agent decision

Neural Computation **33, 483–527** (2021)
https://doi.org/10.1162/neco_a_01346

making at the testing phase. The NE process is applied after the training phase to denoise the memory network so that derived relations are formed in the memory network and retrieved during the testing phase. During the NE phase, indirect relations are enhanced, and the structure of episodic memory changes. This approach can also be interpreted as the agent's replay after the training phase, which is in line with recent findings in behavioral and neuroscience studies. In comparison with EPS, our model is able to model the formation of derived relations and other features such as the nodal effect in a more intrinsic manner. Decision making in the test phase is not an ad hoc computational method, but rather a retrieval and update process of the cached relations from the memory network based on the test trial. In order to study the role of parameters on agent performance, the proposed model is simulated and the results discussed through various experimental settings.

1 Introduction

Stimulus equivalence (SE), a phenomenon that Sidman (1971) identified and explored, refers to the condition that members of an equivalence class evoke the same response in human and animal subjects. The SE methodology uses a matching-to-sample (MTS) procedure to train arbitrary relations between unfamiliar stimuli and test derived relations through mathematical relations in equivalence sets: reflexivity, symmetry, and transitivity. The SE framework, as an efficient learning method, has been widely studied by employing humans or animals as experimental participants (see Sidman, Cresson, & Willson-Morris, 1974; Sidman et al., 1982; Sidman & Tailby, 1982; Sidman, Willson-Morris, & Kirk, 1986; Devany, Hayes, & Nelson, 1986; Hayes, 1989; Fields, Adams, Verhave, & Newman, 1990; Spencer & Chase, 1996; Groskreutz, Karsina, Miguel, & Groskreutz, 2010; Steingrimsdottir & Arntzen, 2011; Arntzen & Mensah, 2020, to mention a few). Computational models constitute another alternative for understanding SE and studying variables that are challenging to examine on humans or animals due to time constraints or ethical issues (see, e.g., Barnes & Hampson, 1993; Cullinan, Barnes, Hampson, & Lyddy, 1994; Lyddy, Barnes-Holmes, & Hampson, 2001; Lew & Zanutto, 2011; Tovar & Westermann, 2017; Ninness, Ninness, Rumph, & Lawson, 2018, for some computational models of the learning of equivalence relations).

In our previous model (Mofrad, Yazidi, Hammer, & Arntzen, 2020), we proposed equivalence projective simulation (EPS) for computationally modeling the SE phenomenon. In brief, EPS has modeled the formation of SE classes through an MTS procedure. A projective simulation (PS) framework (Briegel & De las Cuevas, 2012) was the basis of the model, and we have proposed several methods to address the test phase and derived relations, including max-product, memory sharpness, and random walk on the memory network with absorbing action sets. The EPS model, similar to the original PS model, has an internal episodic memory that is updated during the training phase, which is used to cope with new, derived relations in the testing phase. The PS model, and therefore the EPS model, is flexible and easy to interpret, which allows modeling a broad range of behaviors in MTS experiments, including typical participants or participants with some disabilities. Many parameters of the model can be controlled, such as the learning rate, forgetting rate, and nodal effect.

The EPS model relies on the assumption that the relations are derived on request, that is, when they appear in an MTS trial during the testing phase and updated during the training phase. We slightly change this assumption and form those relations at the end of the training phase; thus, the output network from the training phase of EPS is assumed to be a noisy version of the agent's memory network that is supposed to contain all trained and derived relations. Using a denoising approach, we could produce a new, less noisy clip network that contains information regarding equivalence class formation. The trained relations in the training phase are mapped into a transition matrix whose values describe the strength of the trained relations. By resorting to network enhancement (Wang et al., 2018), we address the formation of SE classes using an iterative update of the transition matrix. Interestingly, the updating process permits naturally denoising the transition matrix and enhancing indirect relations[1] while preserving the initial direct relations learned during the training phase. The denoised network can be assimilated to an updated clip network, used later in the testing phase. It can also be used to assess overall agent performance on eventual equivalence tests. In summary, the contribution of this letter is as follows:

1. Instead of using reasoning, that is, computing the likelihood of the different alternatives during testing by following some indirect paths over the clip network, we update memory and retrieve the updated memory at the testing phase.
2. As in the EPS model, we still control symmetry relations with a multiplicative parameter. We are able to control the ability to derive transitivity relations using parameter α. This turns out to be of great importance when modeling subjects with learning disabilities.
3. We further enhance the NE and propose DNE in which we can control the agent's ability to derive symmetry and also control its ability to derive transitivity.
4. A comparison of PS, EPS, and E-EPS, together with supporting studies from the neuroscience literature, is provided that justifies the proposed model.

[1] According to the theory of SE, indirect relations are derived through reflexivity, symmetry, transitivity, and equivalence.

5. From a computational point of view, the new updating rule has fewer parameters to fine-tune in comparison with the EPS. The approach to deriving relations in EPS model can be seen as routing in the clip network, with action sets as destination points. In the E-EPS model, a diffusion model explores the clip network by simultaneous propagation of flow without a specific target.
6. The updated clip network can be considered as a cognitive map of stimuli that can be used in analyzing the results of different settings.
7. The testing phase in the E-EPS model involves less computation on the decision time in comparison with EPS. E-EPS uses the updated network during the testing phase rather than processing the trained relations to compute derived relation links at each test trial.
8. Using a simulation of several configurations, we study the parameters in detail.
9. We compare three training procedures—linear series (LS), many-to-one (MTO), and one-to-many (OTM)—in the final experiment. In line with the mainstream literature in behavior analysis (see Arntzen, Grondahl, & Eilifsen, 2010; Arntzen & Hansen, 2011; Arntzen, 2012), the model yields better performance in OTM and MTO cases in comparison with LS, which is a qualitative property of our model confirming that it is a realistic model.
10. We provide theoretical analysis of the model and a convergence guarantee in appendix A.

We provide a brief overview of SE, EPS, and network enhancement in section 2. We provide the architecture of the enhanced equivalence projective simulation (E-EPS) model in section 3, where we also compare the proposed approach to the original PS model and recent EPS model. We consider seven experimental scenarios to study the parameters of the model in section 4. Section 5 offers a summary of the letter discussion, and concluding remarks.

2 Background and Related Work

In section 2.1, we review the concept of SE from a behavior analysis perspective. In section 2.2, we briefly explain the EPS model and provide a brief section about network enhancement (Wang et al., 2018) in section 2.3.

2.1 Stimulus Equivalence (SE). SE is a research method on complex human behavior, including memory and problem solving (Sidman, 1990). In the MTS or conditional discrimination procedure, which is used in SE, a given stimulus, say A_1, must be paired with B_1 among a given comparison stimuli set, say B_1, B_2, and B_3. The discrimination happens through programmed consequences.

The MTS procedure has two phases: the training phase, when the participant learns some relations, and the testing phase, when the participant is tested with derived relations. Trial types in the testing phase include baseline, symmetry, transitivity, and equivalence. It is noteworthy that equivalence relations are sometimes referred to as combined transitivity and symmetry.

The evaluation of participant learning is usually through a threshold or mastery criterion ratio (e.g., 0.95 to 1). If the participant passes the criterion, the derived relations are tested. In the testing phase, there are no programmed consequences, and usually the criterion ratio in this phase is lower than in training phase (e.g., 0.9 to 1). Whenever the evidence (passing the criterion for testing) shows the emergence of all relations, the equivalence class is considered to be formed (Sidman & Tailby, 1982).

In the equivalence literature, three training structures have been used in establishing conditional discrimination with the MTS procedure: linear series (LS), many-to-one (MTO), and one-to-many (OTM) (see Arntzen, 2012, for more details about MTS training and testing procedures and parameters in SE formation). Generally, a class with n stimuli, requires training of only $(n-1)$ stimulus-stimulus relations. The condition is that each component of these relations needs to be present in at least one trained relation, and none of the trained relations can have the same two stimuli as components. Even with these constraints, many possible ways for structuring training relations remain, some of them possibly more efficient than the others (see Fields et al., 1990; O'Mara, 1991; Arntzen & Holth, 1997; Hove, 2003; Lyddy & Barnes-Holmes, 2007; Arntzen et al., 2010; Arntzen & Hansen, 2011; Fienup, Wright, & Fields, 2015, for instance). Appendix B formally analyzes the size of the training design space, which is shown to be exhaustive even for a small number of categories and number of classes. Therefore, it is complex to design and run experiments involving human subjects that explore different training and testing scenarios. Computational models, however, could be used for exploring new ideas through simulation. For instance, one could try several configurations and find the optimum scenario according to some design criterion in the computational model before running a real experiment. Moreover, components of the computational model can be easily manipulated, disrupted, impaired, and removed to see the effect of those components on the results. Having more control over the experimental variables, including a controllable environment, is a considerable advantage of these models over real experiments (Barnes & Hampson, 1993; McClelland, 2009; Ninness, Ninness, Rumph, & Lawson, 2018).

2.2 Equivalence Projective Simulation (EPS).

EPS is based on PS, which can be seen as an reinforcement learning (RL) model that can be embodied in an environment, perceive stimuli, execute actions, and learn through trial and error (see, e.g., Briegel & De las Cuevas, 2012; Melnikov, Makmal, Dunjko, & Briegel, 2017, for details of PS model).

The PS agent, and therefore the EPS agent, has an episodic memory that is literally a directed, weighted network of clips, where each clip represents a remembered percept or action (stimulus in EPS). Memory can be described as a probabilistic network of clips, the so-called clip network.[2] The learning in PS is realized by updating weights and structure through adding new clips and new transition links.

The simulation of the MTS procedure via EPS has two phases: the training phase, when the memory network will be formed through trials and guided feedback, and the testing phase, when no new memory clips are created. Although there is no guided feedback in the testing phase, connection weights might be updated. The testing phase is the main part of the model. In Mofrad et al. (2020) three different approaches dealing with the derived relations are discussed: max-product, memory sharpness, and absorbing action sets.

At the beginning of an MTS training phase, the agent memory space, which is shown by $\mathcal{C} = \{c_1, \ldots, c_p\}$, is empty. Based on trial settings, a memorized clip could play the role of either a percept clip or an action clip. At each time step, the environment (the experimenter in the real experiments) shows a sample stimulus and some comparison alternatives, which are referred to as percept and actions. The percept and actions belong to the percept set $\mathcal{S}$ and action set $\mathcal{A}$, respectively. The sample stimulus (percept, $s \in \mathcal{S}$) and the comparison stimuli (actions $a \in \mathcal{A}_t$) belong to different categories (e.g., category A or B), where $\mathcal{A}_t$ denotes the action space at time t and consists of a set of comparisons at the given trial. The training phase will be as follows:

1. The agent perceives stimulus $s \in \mathcal{S}$ from the environment. Clip $c_s \in \mathcal{C}$ is either created (the first time) or activated.
2. Perceiving action set $\mathcal{A}_t$ from the environment, the agent establishes and initializes connections between the sample and comparison stimuli the first time with h-values equal to h_0. If there exist connections from previous trials, there is no need for initialization.
3. The agent computes $p^{(t)}(c_a|c_s)$, $a \in \mathcal{A}_t$ based on the h-values using the softmax distribution function,

$$p^{(t)}(c_j|c_i) = \frac{e^{\beta h^{(t)}(c_i,c_j)}}{\sum_k e^{\beta h^{(t)}(c_i,c_k)}}, \tag{2.1}$$

where at this stage, clip $c_i = c_s$ and clip $c_j \in \mathcal{A}_t$. A larger value of $\beta \geq 0$ creates a probability distribution that is more biased to the choice of

[2] The terms *episode* and *clip* are used interchangeably.

the largest h-value, and therefore parameter β can be used for tuning the learning rate as well.

4. The agent selects one of the actions based on the computed probability distribution and receives a positive or negative reward from the environment, say, $\lambda^{(t)} \in \Lambda = \{-1, 1\}$.[3]
5. The connection weights, h-values, will be updated as a result of the environment feedback as follows:

$$h^{(t+1)}(c_s, c_a) = h^{(t)}(c_s, c_a) - \gamma(h^{(t)}(c_s, c_a) - 1) + \lambda^{(t)}. \tag{2.2}$$

Moreover, the opposite link, (c_a, c_s), will be updated in a similar way, but with the parameter $0 < K \leq 1$:

$$h^{(t+1)}(c_a, c_s) = h^{(t)}(c_a, c_s) - \gamma(h^{(t)}(c_a, c_s) - 1) + K\lambda^{(t)}. \tag{2.3}$$

6. The environment provides new trials until all training relations meet the mastery criterion.

It is noteworthy that parameter K was used in the learning rule of the original PS model (Briegel & De las Cuevas, 2012) to determine the growth rate of associative or compositional connections relative to the direct connections. This parameter, for instance, enables the PS agent to learn faster by recognizing similarity among the existing clips in memory and new perceptual input (see Figures 11 and 12 in Briegel & De las Cuevas, 2012, for more detail on associative learning in the PS agent). The parameter K in the EPS model, however, quantifies the relative growth of symmetry relations compared to the direct, or baseline, relations.[4] This parameter is different from the original PS in the sense that the stimuli in EPS (and E-EPS) are arbitrary, that is, they have no physical similarity, and therefore the parameter K does not capture similarity. The notion of associative memory, however, can be added to the EPS model by introducing compound stimuli, which we do not address in this letter.

After that agent passes the training phase, the testing phase, in which the formation of derived relations is tested starts. At this stage, no feedback is provided from the environment.

1. The agent perceives $s \in \mathcal{S}$, activates the memory clip $c_s \in \mathcal{C}$, and tries to choose the best action among the given action set $\mathcal{A}_t$ based on its memory as follows.

[3] It is noteworthy that Λ could have any positive or negative values, including asymmetric rewards. For instance, negative feedback might have greater impact (see Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001, as an example of a positive-negative asymmetry effect).

[4] In Mofrad et al. (2020), we use K_1, K_2, K_3, and K_4, which play the same role as K in this letter but with a higher level of control.

2. If connections between the sample and comparisons exist, the agent computes the $p^{(t)}(c_a|c_s), a \in \mathcal{A}_t$ based on the h-values using a probabilistic distribution achieved by either softmax or a normalized vector of h-values (called "standard" in PS and EPS). If such connections do not exist in the transitivity or equivalence relation cases, the agent computes the transition probabilities using a max-product scenario or an absorbing states scenario and selects one of the possible actions:
   - In the max-product case, the agent finds the most probable paths between c_s and each action $c_a, a \in \mathcal{A}_t$. There are many possible paths that might link c_s to a particular action c_a, and thus the procedure might be computationally exhaustive.
   - The absorbing state scenario can be considered as a random clip network, starting from c_s and ending with a clip in $\mathcal{A}_t$. So, unlike the max-product method, the probability of reaching each action from c_s is important but not the path itself. These probabilities can be computed when actions $c_a \in \mathcal{A}_t$ are set to be absorbing states of the underlying Markov chain at time t.
3. Memory sharpness, $0 \leq \theta \leq 1$, functions as a mechanism to control the formation of transitivity relations and consequently controls equivalence relations and the effect of the nodal number (see, e.g., Sidman, 1994, for nodal number), in line with the baseline relations training. Mofrad et al. (2020) discuss memory sharpness as a separate method. However, it can be used in combination with either max-product or the concept of absorbing states.

For the sake of brevity, we review just the parts of the EPS model that are necessary for understanding the new perspective on derived relations. Moreover, an overview of some other behavior-analytic computational approaches to the formation of SE classes is provided in Mofrad et al. (2020), which provides a detailed version of EPS model.

2.3 Network Enhancement (NE). Wang et al. (2018) proposed network enhancement (NE), a computational approach for denoising biological networks. NE converts a noisy, undirected, weighted network into a new network possessing the same nodes but with different connections and weights. It assumes that nodes that are connected through paths with high weight edges have a high chance of being directly connected with a high-weight edge. The NE diffusion process uses random walks of length 3 or less and a regularized information flow in order to produce new edge weights.

For a formal description of NE, let W be the matrix of edge weights and $\mathcal{N}_i$ be the K-nearest neighbors of the ith node, including node i itself. The localized network $\mathcal{T}$ is constructed from W as follows:

$$P_{i,j} \leftarrow \frac{W_{i,j}}{\sum_{k\in\mathcal{N}_i} W_{i,k}} \mathbb{I}_{\{j\in\mathcal{N}_i\}}, \ \mathcal{T}_{i,j} \leftarrow \sum_{k=1}^{n} \frac{P_{i,k}P_{j,k}}{\sum_{v=1}^{n} P_{v,k}}, \tag{2.4}$$

where $\mathbb{I}_{\{.\}}$ is the indicator function. Then the diffusion process is defined as an iterative relation,

$$W_{t+1} = \alpha\mathcal{T} \times W_t \times \mathcal{T} + (1-\alpha)\mathcal{T}, \tag{2.5}$$

where α is a regularization parameter, t shows the iteration step, and W_0 can be initialized with the input matrix W. The update rule in equation 2.5 for each entry is

$$(W_{t+1})_{i,j} = \alpha \sum_{k\in\mathcal{N}_i} \sum_{l\in\mathcal{N}_j} \mathcal{T}_{i,k}(W_t)_{k,l}\mathcal{T}_{l,j} + (1-\alpha)\mathcal{T}_{i,j}. \tag{2.6}$$

The many theoretical properties for this diffusion process are discussed in Wang et al. (2018). It is shown that W_t remains a symmetric, doubly stochastic matrix (DSM) for each iteration t, and W_t converges to a nontrivial equilibrium network. Moreover, NE does not change eigenvectors of the initial DSM $\mathcal{T}$, but the spectrum of the eigenvalues is changed nonlinearly so that the eigengap is increased. This effect of the NE process on the eigenspectrum improves the network to achieve a more accurate detection of clusters. Although this method produces promising results in our model, as we will explain in section 4, it is not the main approach for formation of equivalence classes in the EPS model, but NE and discussions in Wang et al. (2018) are the main motivation for the update rule. The method we use does not have all the properties that NE has, and we refer to the theoretical aspect of the diffusion process we used in appendix A. In the rest of this letter, we refer to the NE method due to Wang et al. (2018) as *symmetric network enhancement* (SNE).

3 Enhanced Equivalence Projective Simulation (E-EPS)

The training phase of the proposed E-EPS model is generally the same as the original PS and the EPS in the sense that the clip network is formed by adding new clips and updating the *h*-values based on the environment feedback. However, since in this letter, the probability distribution over the action set is modeled using the softmax function, we let the network have negative *h*-values and simplify the training by removing some parameters associated with positive *h*-values. However, the approach to the formation of SE classes and the testing phase is quite different compared to the EPS model (Mofrad et al., 2020). As we explained in section 2.2, after the training phase, we have a network of *h*-values for baseline relations and the symmetry relations. To add reflexivity to the clip network, we can consider an updating method either during the training phase[5] or after the training phase. In order to keep the model simple, we add a self-loop to each clip after the training phase and assign it an h-value equal to the maximum h-value of input or output connections. The argument is that in the case where the agent can identify the members of a class (say, A_1, B_1, C_1), it must be able to differentiate members of each category (say, A_1 from A_2 and A_3). We refer to the adjacency matrix of this network of h-values as W_h.

In this work, we are proposing a new NE model called *directed network enhancement* (DNE) that can be used for the testing phase, including baseline, reflexivity, symmetry, transitivity, and equivalence relations. Consider the following rule as the update rule (or diffusion process),

$$W_{t+1} = \alpha P \times W_t \times P + (1-\alpha)P, \tag{3.1}$$

where W_0 is a right stochastic matrix achieved from W_h. (By a "right stochastic matrix," we mean a real square matrix in which each row sums to one.) We put $W_0 = P$ where P is the transition probability matrix of W_h applying softmax function on nonzero values at each row using β_h parameter. P is not symmetric, and $P\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ represents the all-one eigenvector of P associated with eigenvalue one. In other words, P is a right stochastic matrix, so it can be used as the initial matrix in the DNE process. In the theoretical analysis of the SNE process provided by Wang et al. (2018), and supplementary note 3, the converged network is proved to be

$$W_{t\to\infty} = (1-\alpha)\mathcal{T}(\mathcal{I} - \alpha\mathcal{T}^2)^{-1}. \tag{3.2}$$

As we discuss in appendix A, the convergence in the DNE process remains valid for a network where we substitute $\mathcal{T}$ with P in equation 3.2:

$$W_{t\to\infty} = (1-\alpha)P(\mathcal{I} - \alpha P^2)^{-1}. \tag{3.3}$$

This post-processing phase transforms the h-value network obtained by training into a new network that can represent the agent predictive representations in a cognitive map (or successor representation similar to Momennejad, Russek et al., 2017).

The $W_{t\to\infty}$ matrix can be seen as the memory representation where we ignore the effect of context (or actions) and assume all the transitions in the network are based on the random walk on the graph (or diffusion). For instance, we can interpret the $(i,\ j)$ entry of the $W_{t\to\infty}$ matrix as the transition probability from clip i to clip j when there is no external control.

[5] For instance, this can simply be achieved by adding a self-loop edge initialized with h_0 to each clip the first time it is perceived by the agent and update it whenever it gets involved in a trial.

When it comes to the testing phase, the softmax function with β_t is applied to calculate the probability distribution for each test trial. In order to accommodate the controlling effect of the test trials, the input values to the softmax function are set to be conditional probabilities given the trial, which can be calculated using Bayes' rule. As an example, if the test trial consists of A_1 as the sample stimulus and $F = \{F_1, F_2, F_3\}$ as the comparison stimuli, input values for the softmax function are $P(A_1F_1|A_1F), P(A_1F_2|A_1F)$, and $P(A_1F_3|A_1F)$ where event A_1F is either $A_1F_1, A_1F_2,$ or A_1F_3. These conditional values can be calculated due to Bayes' rule, for instance,

$$\begin{aligned}
&P(A_1F_1|A_1F)\\
&\quad= \frac{P(A_1F_1)P(A_1F|A_1F_1)}{P(A_1F_1)P(A_1F|A_1F_1) + P(A_1F_2)P(A_1F|A_1F_2) + P(A_1F_3)P(A_1F|A_1F_3)}\\
&\quad= \frac{P(A_1F_1)}{P(A_1F_1) + P(A_1F_2) + P(A_1F_3)},
\end{aligned}$$

which can be seen as a normalization. Note that all the conditional probabilities on the right-hand side are equal to one and therefore are removed. Parameter β_t in the softmax function can characterize the agent's memory and ability to link an internal representation to the real action. When a test trial is given to the agent, the memory is conditioned based on the test trials (sample and comparison stimuli), and Bayes' rule is used to characterize the environment effect.

Another way to formalize the behavior of the agent in the testing phase is to use a trial-based β_t for the softmax function, which is defined as β_t divided by the summation over weights for comparison stimuli. The above example, with A_1 as the sample stimulus and $F = \{F_1, F_2, F_3\}$ as the comparison stimuli, uses $\frac{\beta_t}{P(A_1F_1)+P(A_1F_2)+P(A_1F_3)}$ as the β in softmax function. As is clear from the example, in this formalization, the results remain exactly the same but they open up room to interpret the agent behavior differently. Using Bayes's rule and a fixed β_t approach emphasizes the effect of environment and the agent characteristics separately, but the variable β_t approach avoids the interpretation that the agent probabilities are calculated twice.

Before comparing the E-EPS to the original PS and the EPS model and relating it to other studies, we summarize the parameters of the agent model:

1. Parameter $0 < K \leq 1$ controls the formation of symmetry relations. $K = 1$ means that the relations are bidirectional and the h-value network is symmetric (see experiment 2).
2. Parameter $0 \leq \gamma < 1$ represents the forgetting rate during the training phase. The training structure (order of relations to be trained) is more important when the forgetting rate is high (see experiment 4).

3. Parameter $\beta_h > 0$ converts h-values to probabilities during the training trials and generates the input matrix W_0 for the NE process (see experiments 1 and 3).
4. Parameter $0 \leq \alpha < 1$ controls to what extent the NE affects the initially trained network when there is no test trial in place. α could characterize the amount of abstract mental process or replay that the agent performs. Even a small value of α could form derived relations that are weak compared to direct relations, but the ratio or conditional probabilities (used as an input to the softmax function) are strong. A value close to one for α means too much diffusion, which can erase the trained relations. One might find the appropriate diffusion based on the expected agent abilities and the training criterion (see experiment 5 and appendix A for more details)
5. Parameter $\beta_t > 0$ controls the agent's performance in a test trial (see experiment 6).

3.1 PS, EPS, and E-EPS: Discussion and Comparison. As Briegel and De las Cuevas (2012) mentioned, the idea of a clip network in PS is similar to the idea of Tolman's (1948) cognitive maps, which refers to a rich internal model of the world that represents relationships between events and simulates the consequences of actions. Although cognitive maps are mostly used for modeling spatial behavior (O'Keefe & Nadel, 1978), they are more general and cover the organization of knowledge in other types of behaviors, including flexible behavior. Cognitive maps can be constructed from abstract representations to describe relational knowledge, and new cognitive problems can then be considered as inference in this relational basis (Behrens et al., 2018).

Brain studies suggest multiple solutions to predicting long-term reward in RL problems (Daw, Niv, & Dayan, 2005). Learning a model of the environment, or a cognitive map of the environment, and using it to simulate future states step-by-step to predict long-term reward are different solutions, which we refer to as model-based RL (Daw et al., 2005; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Sutton & Barto, 2018). Forming simple world models in the human hippocampus for relational knowledge sorting and value spreading across associated stimulus representations is shown to directly influence behavior in a novel decision-making situations (Wimmer & Shohamy, 2012). Repeating patterns during both awake experiential states and nonengaged states and reshaping neural circuits has been studied in both the hippocampus and the neocortex (see Liu & Watson, 2020, for a review). Functional magnetic resonance imaging (fMRI) similarity measures in the hippocampus and entorhinal cortex (Stachenfeld, Botvinick, & Gershman, 2017; Garvert, Dolan, & Behrens, 2017) suggest the existence of statistical transitions of discrete state-spaces. The use of precompiled transition distances, rather than simulating all possible transitions online, is studied by Momennejad, Russek et al. (2017), where these precompiled distances depend on offline activity, or replay, in the hippocampus and ventral frontal cortex (Momennejad, Otto, Daw, & Norman, 2017). Caching of multistep predictive representations is also referred to as a “predictive map” (Stachenfeld et al., 2017). These predictive representations link model-based RL to model-free mechanisms through offline replay mechanisms (Russek, Momennejad, Botvinick, Gershman, & Daw, 2017) resembling Dyna-style planning (Sutton, Szepesvári, Geramifard, & Bowling, 2008).

PS is much more primitive than Dyna-style planning. It only changes the weights of the clip transition and performs a random walk on the clip network (for a detailed comparison, see Mautner, Makmal, Manzano, Tiersch, & Briegel, 2015). The multiple reflection in the PS model is different from “experiment replay” (Lin, 1992) in the sense that PS uses short-term memory, or emotional tags, to evaluate the result of a simulation and repeat the random walk if the remembered reward for the chosen action in the previous round was negative. So repeatedly presenting its experiences to its learning algorithm is not performed just for the sake of memory consolidation. (See also Momennejad, 2020, for a review on the role of replay on how the brain learns and generalizes relational structures with a focus on the RL approach.)

In the EPS model (Mofrad et al., 2020), two scenarios, called “standard” and “softmax,” were used for the training phase, and various ways for deriving relations in the test phase were studied and discussed due to the aim to define EPS as a general and flexible model. The EPS (and E-EPS) training phase is similar to the PS model with extra links and update rule for symmetry relations. In this letter, we survey just the training method that uses the softmax function in order to calculate probability distributions over the action sets. Although the training phase in this letter could be similar to EPS, for simplicity, we just consider the softmax scenario where negative *h*-values are allowed, so we can formalize the learning with just one parameter, *K*, to control the growth ratio of symmetry relations in comparison with the direct relations.

The main difference with PS, the most important part of the EPS (E-EPS) model, is the testing phase where there is no feedback. In the EPS model, the derived relations were calculated on demand at the decision time whenever they appear in a test trial. The probabilities are either calculated based on the probabilities of the paths with maximum values, using a max-product algorithm, or the probability of reaching each of the action points having a random walk on the episodic memory started at a sample stimulus. The symmetry relations are controlled via a multiplicative parameter, and the transitivity can be controlled with a parameter called memory sharpness.

In the EPS testing phase, the only change to the clip network *h*-values is related to the parameter γ, the forgetting factor, and all the computations for the test trials are performed at the decision time, which can be seen as an ad hoc computational tool rather than an intrinsic feature of the model. The perspective to the derived relation in E-EPS, is quite different where NE, an iterative diffusion process, is used after the training phase. This alternative approach updates the structure of clip network by adding new connections between the clips and updating connection weights. In other words, the approach to derive relations in the EPS model can be seen as routing in the clip network, where the action sets play the role of destination, while in the E-EPS model, in the absence of test trials, the approach involves a diffusion model to explore the clip network by simultaneous propagation of flow without a specific target. The NE process is in line with the random walk-based decision making in the PS approach. It is noteworthy that diffusion models have been successfully used in various cognitive tasks involving decision making (Shrager, Hogg, & Huberman, 1987; Ratcliff, Smith, Brown, & McKoon, 2016). Stella et al. (2019) show that hippocampal circuits can reactivate random trajectories of varying lengths and timescales that resemble Brownian diffusion. The NE process can also be interpreted as a kind of replay similar to the offline replay that contributes to generalization via multistep predictive representations of upcoming clips (or the successor representation; see Momennejad, Otto et al., 2017; Momennejad, Russek et al., 2017; Russek et al., 2017. It is different from online replay or multiple reflection in the PS model and closer to the offline replay that accommodates planning based on inferential piecing data together and multistep dependencies. The REMERGE (recurrency and episodic memory results in generalization) model of memory trace activation (Kumaran & McClelland, 2012) also uses replay and iterative updating of episodic memory for modeling rapid generalization in, for example, transitive inference task.

The final abilities of the E-EPS agent to master derived relations strongly depends on two parameters: α, which controls how much the NE affects the initially trained network, and β_t, which generates the probability distribution over the comparison stimuli. The post-processed network, $W_{t\to\infty}$, can be seen as an unconditioned network that a test trial can bias it. To account for the environment effect, we use a Bayesian approach and then apply the softmax function (see McClelland, 2013, for different models of contextual effects on perception). It is noteworthy that the PS model uses Bayesian updating, and therefore this update is in harmony with the PS agent (see Schwöbel, Kiebel, & Marković, 2018, and Parr, Markovic, Kiebel, & Friston, 2019, for modeling goal-directed behavior as an inference process).

The approach to the testing phase in the E-EPS model needs less computation at the decision time since it uses the cached updated network rather than processing the trained relations to compute derived relation links at each test trial.

In the rest of the letter, we discuss and conduct experiments on both models SNE and DNE, but the emphasis will be on the DNE, which we show is more effective than the SNE for the E-EPS model.

Table 1: Training Stages in Spencer and Chase (1996) Study: Number and Type of Training Trials.

	Number of Trials per Relation					
Training	*AB*	*BC*	*CD*	*DE*	*EF*	*FG*
AB	48					
BC	24	24				
CD	12	12	24			
DE	9	9	9	24		
EF	6	6	6	6	24	
FG	3	3	3	6	9	24
Baseline maintenance	3	3	3	3	3	3

4 Simulation Results

In this section, we study the model parameters in order to offer insights into how parameters can be tuned to simulate various behaviors, including typical human behavior or the behavior of people with some disabilities. To study the model in more detail, we consider a similar training setting as in the experiment by Spencer and Chase (1996), which Mofrad et al. (2020) address as well.

Spencer and Chase (1996) address the relatedness or nodal number using three seven-member stimulus classes. Stimuli are nonsense figures, and the training order is $A \to B \to C \to D \to E \to F \to G$. The training consists of seven stages as summarized in Table 1.[6] The first training block contains 48 trials of *AB* relations. Since there are three classes, the block for training, *AB*, contains 16 trials with the correct match A_1B_1, 16 trials with the correct match A_2B_2, and 16 trials with the correct match A_3B_3. The order of presented trials is random in the block, and the order of comparison stimuli, in this case B_1, B_2, B_3, is also randomly changed. If we consider the training of the *EF* relation, for instance, the training block contains six *AB* relations (which means each trial with A_1B_1, A_2B_2, and A_3B_3 as the correct pair appears twice), 6 *BC* relations (each trial with B_1C_1, B_2C_2, and B_3C_3 as the correct pair appears twice), 6 *CD* relations and 6 *DE* relations, and finally the new relation *EF* with 24 relations (i.e., each trial with E_1F_1, E_2F_2, and E_3F_3 as the correct pair appears eight times). In the baseline maintenance stage, no new relation is provided and each correct relation appears only once. The

[6] It is noteworthy that in Spencer and Chase (1996), each stage of training has 48 trials. To ease the simulation, the fourth stage for *DE* training is changed, so we consider 9 instead of 8 trials for *AB*, *BC*, and *CD* relations. Therefore, this stage has 51 trials in the simulation instead of the original 48.

mastery criterion is set to 0.9, and if the agent passes the mastery criterion for all stages and the final baseline maintenance, then we can test the agent for formation of derived relations.

The reported simulation results are the average over 1000 simulations.

4.1 Experiment 1: Step-by-Step Process. In this experiment, we illustrate the computation steps. In Figure 1a, the network h-values after the training phase (based on Table 1) is depicted where the parameters are set to $\gamma = 0.001$, $K = 1$, $\beta_h = 0.1$, $\beta_t = 4$, and $\alpha = 0.7$. Note that the symmetry and reflexivity connections in addition to the baseline connections appear in Figure 1a. The reflexivity h-values are the maximum h-value at each row (input-output connections). Moreover, since $K = 1$, the W_h matrix is symmetric—for instance, $A_1B_1 = B_1A_1 = 51.82$. To compute the transition probability matrix, a softmax function with parameter $\beta_h = 0.1$ is used. Note that the transition probability matrix is just row-normalized and not symmetric. All the reported values are rounded by two or three decimal places.

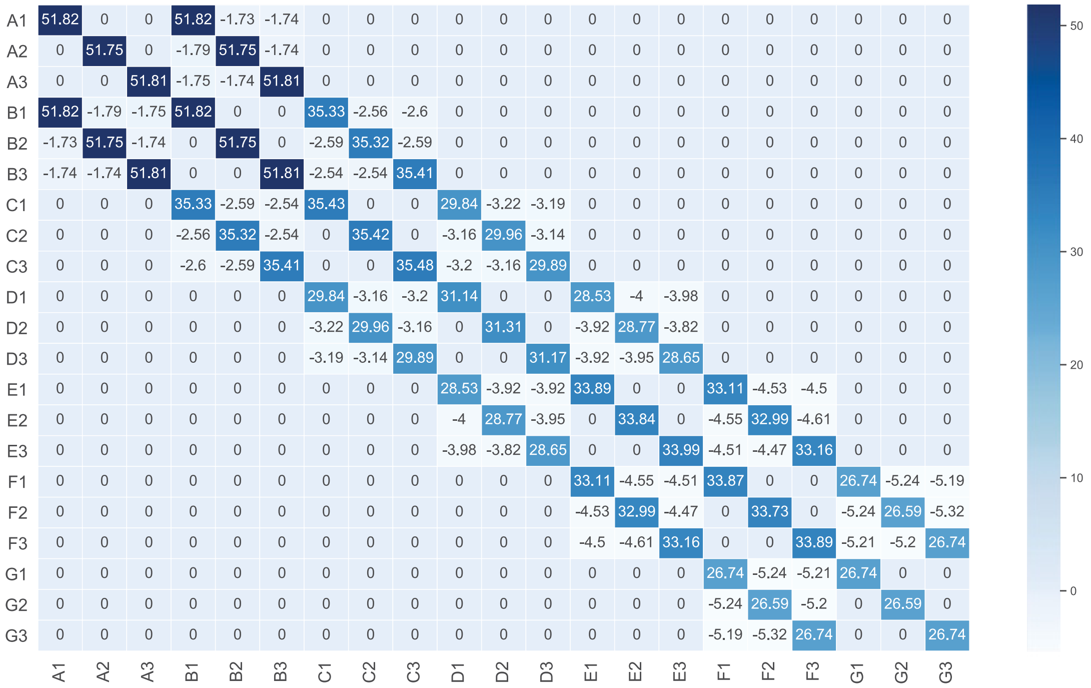
We set $W_0 = P$ as the input matrix to the NE. We might use P (Figure 1b) for the iterative updates (DNE) or $\mathcal{T}$ matrix (SNE). In Figure 2, we address DNE when $\alpha = 0.7$. The convergence criterion is that $\sum_{i,j} |W_{t+1} - W_t|_{i,j} < 0.0001$. One can also compute the converged network $W_{t\to\infty}$ using the theoretical converged formula provided in equation 3.3.

Figure 2a shows the general internal map of the network clip before the testing phase. One can interpret these values as how the stimuli are prioritized in the agent memory when there is no external trial that measures the accuracy of answers in MTS trials. Figure 2b shows the performance of the agent when it comes to the testing phase. For instance, if the sample stimulus is A_3 and the comparison stimuli are F_1, F_2, and F_3, then the agent chooses F_1 and F_2 with probability 0.092 and selects F_3 with probability 0.815.

To calculate these category-based probability distributions, first the conditional probability for any specific category is calculated based on Bayes' rule, and then the softmax function transfers these vectors to the desired probabilities based on the chosen parameter β_t. The conditional input can show the context, or environment, effect, and therefore we can apply the same β_t as a characteristic of the agent for all the categories.

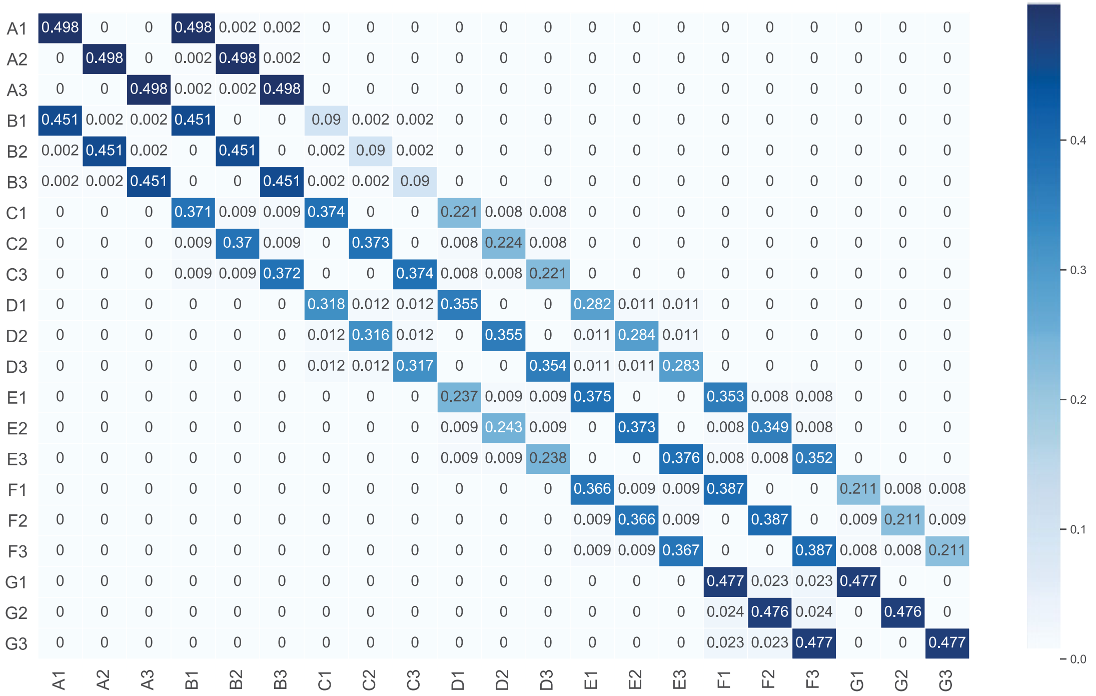
If we use SNE, first we have to compute $\mathcal{T}$, which is reported in Figure 3a and then update the network using $\alpha = 0.7$ parameter. The localized network $\mathcal{T}$ adds weights to the one-node relations, and we have two more diagonals in $\mathcal{T}$ in comparison with P.

The goal of this experiment is to illustrate how both DNE and SNE are working. In experiment 2, we compare the two updating methods for symmetry and transitivity relations and discuss why DNE could be a more appropriate option for enhancing the EPS model.



	A1	A2	A3	B1	B2	B3	C1	C2	C3	D1	D2	D3	E1	E2	E3	F1	F2	F3	G1	G2	G3
A1	51.82	0	0	51.82	-1.73	-1.74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A2	0	51.75	0	-1.79	51.75	-1.74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A3	0	0	51.81	-1.75	-1.74	51.81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1	51.82	-1.79	-1.75	51.82	0	0	35.33	-2.56	-2.6	0	0	0	0	0	0	0	0	0	0	0	0
B2	-1.73	51.75	-1.74	0	51.75	0	-2.59	35.32	-2.59	0	0	0	0	0	0	0	0	0	0	0	0
B3	-1.74	-1.74	51.81	0	0	51.81	-2.54	-2.54	35.41	0	0	0	0	0	0	0	0	0	0	0	0
C1	0	0	0	35.33	-2.59	-2.54	35.43	0	0	29.84	-3.22	-3.19	0	0	0	0	0	0	0	0	0
C2	0	0	0	-2.56	35.32	-2.54	0	35.42	0	-3.16	29.96	-3.14	0	0	0	0	0	0	0	0	0
C3	0	0	0	-2.6	-2.59	35.41	0	0	35.48	-3.2	-3.16	29.89	0	0	0	0	0	0	0	0	0
D1	0	0	0	0	0	0	29.84	-3.16	-3.2	31.14	0	0	28.53	-4	-3.98	0	0	0	0	0	0
D2	0	0	0	0	0	0	-3.22	29.96	-3.16	0	31.31	0	-3.92	28.77	-3.82	0	0	0	0	0	0
D3	0	0	0	0	0	0	-3.19	-3.14	29.89	0	0	31.17	-3.92	-3.95	28.65	0	0	0	0	0	0
E1	0	0	0	0	0	0	0	0	0	28.53	-3.92	-3.92	33.89	0	0	33.11	-4.53	-4.5	0	0	0
E2	0	0	0	0	0	0	0	0	0	-4	28.77	-3.95	0	33.84	0	-4.55	32.99	-4.61	0	0	0
E3	0	0	0	0	0	0	0	0	0	-3.98	-3.82	28.65	0	0	33.99	-4.51	-4.47	33.16	0	0	0
F1	0	0	0	0	0	0	0	0	0	0	0	0	33.11	-4.55	-4.51	33.87	0	0	26.74	-5.24	-5.19
F2	0	0	0	0	0	0	0	0	0	0	0	0	-4.53	32.99	-4.47	0	33.73	0	-5.24	26.59	-5.32
F3	0	0	0	0	0	0	0	0	0	0	0	0	-4.5	-4.61	33.16	0	0	33.89	-5.21	-5.2	26.74
G1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26.74	-5.24	-5.21	26.74	0	0
G2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-5.24	26.59	-5.2	0	26.59	0
G3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-5.19	-5.32	26.74	0	0	26.74

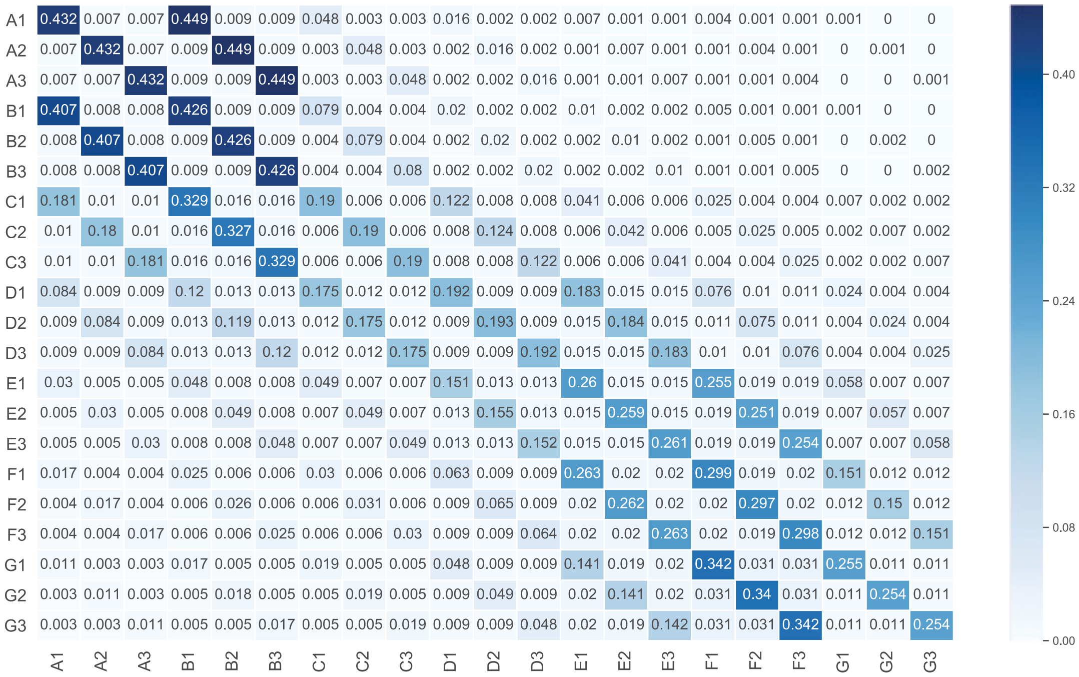
(a) Network clip W_h, composed of h-values at the end of training phase.



	A1	A2	A3	B1	B2	B3	C1	C2	C3	D1	D2	D3	E1	E2	E3	F1	F2	F3	G1	G2	G3
A1	0.498	0	0	0.498	0.002	0.002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A2	0	0.498	0	0.002	0.498	0.002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A3	0	0	0.498	0.002	0.002	0.498	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1	0.451	0.002	0.002	0.451	0	0	0.09	0.002	0.002	0	0	0	0	0	0	0	0	0	0	0	0
B2	0.002	0.451	0.002	0	0.451	0	0.002	0.09	0.002	0	0	0	0	0	0	0	0	0	0	0	0
B3	0.002	0.002	0.451	0	0	0.451	0.002	0.002	0.09	0	0	0	0	0	0	0	0	0	0	0	0
C1	0	0	0	0.371	0.009	0.009	0.374	0	0	0.221	0.008	0.008	0	0	0	0	0	0	0	0	0
C2	0	0	0	0.009	0.37	0.009	0	0.373	0	0.008	0.224	0.008	0	0	0	0	0	0	0	0	0
C3	0	0	0	0.009	0.009	0.372	0	0	0.374	0.008	0.008	0.221	0	0	0	0	0	0	0	0	0
D1	0	0	0	0	0	0	0.318	0.012	0.012	0.355	0	0	0.282	0.011	0.011	0	0	0	0	0	0
D2	0	0	0	0	0	0	0.012	0.316	0.012	0	0.355	0	0.011	0.284	0.011	0	0	0	0	0	0
D3	0	0	0	0	0	0	0.012	0.012	0.317	0	0	0.354	0.011	0.011	0.283	0	0	0	0	0	0
E1	0	0	0	0	0	0	0	0	0	0.237	0.009	0.009	0.375	0	0	0.353	0.008	0.008	0	0	0
E2	0	0	0	0	0	0	0	0	0	0.009	0.243	0.009	0	0.373	0	0.008	0.349	0.008	0	0	0
E3	0	0	0	0	0	0	0	0	0	0.009	0.009	0.238	0	0	0.376	0.008	0.008	0.352	0	0	0
F1	0	0	0	0	0	0	0	0	0	0	0	0	0.366	0.009	0.009	0.387	0	0	0.211	0.008	0.008
F2	0	0	0	0	0	0	0	0	0	0	0	0	0.009	0.366	0.009	0	0.387	0	0.009	0.211	0.009
F3	0	0	0	0	0	0	0	0	0	0	0	0	0.009	0.009	0.367	0	0	0.387	0.008	0.008	0.211
G1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.477	0.023	0.023	0.477	0	0
G2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.024	0.476	0.024	0	0.476	0
G3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.023	0.023	0.477	0	0	0.477

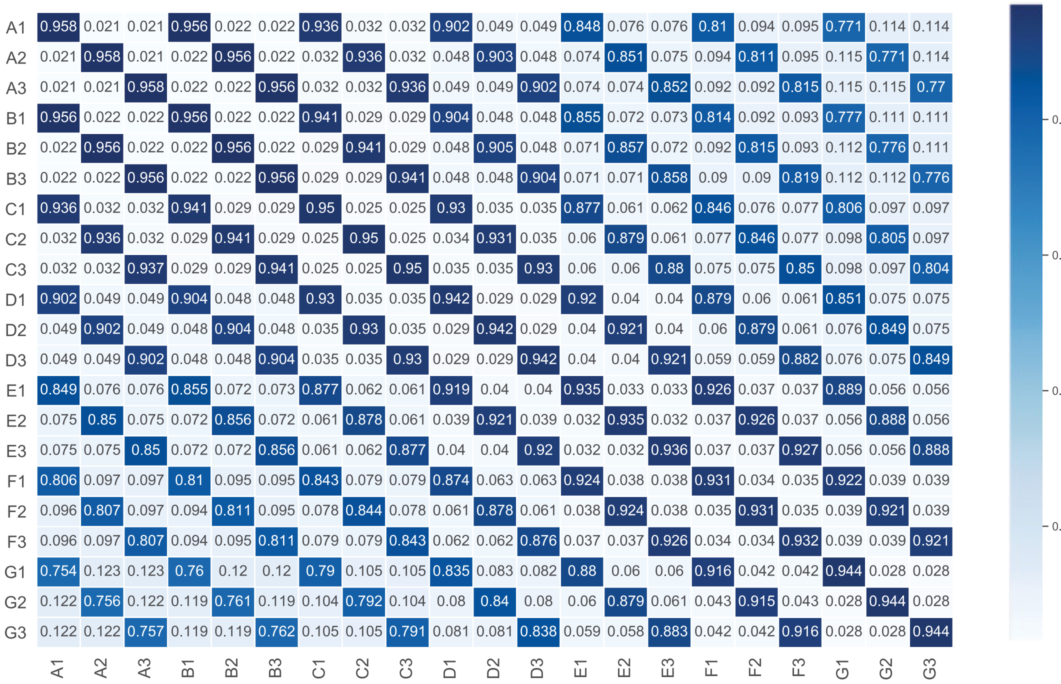
(b) The transition probability matrix P using $\beta_h = 0.1$. The reported values are rounded by three decimal places.

Figure 1: A sample configuration of network h-values after training $A \to B \to C \to D \to E \to F \to G$ when $\gamma = 0.001$, $K = 1$, and $\beta_h = 0.1$.



	A1	A2	A3	B1	B2	B3	C1	C2	C3	D1	D2	D3	E1	E2	E3	F1	F2	F3	G1	G2	G3
A1	0.432	0.007	0.007	0.449	0.009	0.009	0.048	0.003	0.003	0.016	0.002	0.002	0.007	0.001	0.001	0.004	0.001	0.001	0.001	0	0
A2	0.007	0.432	0.007	0.009	0.449	0.009	0.003	0.048	0.003	0.002	0.016	0.002	0.001	0.007	0.001	0.001	0.004	0.001	0	0.001	0
A3	0.007	0.007	0.432	0.009	0.009	0.449	0.003	0.003	0.048	0.002	0.002	0.016	0.001	0.001	0.007	0.001	0.001	0.004	0	0	0.001
B1	0.407	0.008	0.008	0.426	0.009	0.009	0.079	0.004	0.004	0.02	0.002	0.002	0.01	0.002	0.002	0.005	0.001	0.001	0.001	0	0
B2	0.008	0.407	0.008	0.009	0.426	0.009	0.004	0.079	0.004	0.002	0.02	0.002	0.002	0.01	0.002	0.001	0.005	0.001	0	0.002	0
B3	0.008	0.008	0.407	0.009	0.009	0.426	0.004	0.004	0.08	0.002	0.002	0.02	0.002	0.002	0.01	0.001	0.001	0.005	0	0	0.002
C1	0.181	0.01	0.01	0.329	0.016	0.016	0.19	0.006	0.006	0.122	0.008	0.008	0.041	0.006	0.006	0.025	0.004	0.004	0.007	0.002	0.002
C2	0.01	0.18	0.01	0.016	0.327	0.016	0.006	0.19	0.006	0.008	0.124	0.008	0.006	0.042	0.006	0.005	0.025	0.005	0.002	0.007	0.002
C3	0.01	0.01	0.181	0.016	0.016	0.329	0.006	0.006	0.19	0.008	0.008	0.122	0.006	0.006	0.041	0.004	0.004	0.025	0.002	0.002	0.007
D1	0.084	0.009	0.009	0.12	0.013	0.013	0.175	0.012	0.012	0.192	0.009	0.009	0.183	0.015	0.015	0.076	0.01	0.011	0.024	0.004	0.004
D2	0.009	0.084	0.009	0.013	0.119	0.013	0.012	0.175	0.012	0.009	0.193	0.009	0.015	0.184	0.015	0.011	0.075	0.011	0.004	0.024	0.004
D3	0.009	0.009	0.084	0.013	0.013	0.12	0.012	0.012	0.175	0.009	0.009	0.192	0.015	0.015	0.183	0.01	0.01	0.076	0.004	0.004	0.025
E1	0.03	0.005	0.005	0.048	0.008	0.008	0.049	0.007	0.007	0.151	0.013	0.013	0.26	0.015	0.015	0.255	0.019	0.019	0.058	0.007	0.007
E2	0.005	0.03	0.005	0.008	0.049	0.008	0.007	0.049	0.007	0.013	0.155	0.013	0.015	0.259	0.015	0.019	0.251	0.019	0.007	0.057	0.007
E3	0.005	0.005	0.03	0.008	0.008	0.048	0.007	0.007	0.049	0.013	0.013	0.152	0.015	0.015	0.261	0.019	0.019	0.254	0.007	0.007	0.058
F1	0.017	0.004	0.004	0.025	0.006	0.006	0.03	0.006	0.006	0.063	0.009	0.009	0.263	0.02	0.02	0.299	0.019	0.02	0.151	0.012	0.012
F2	0.004	0.017	0.004	0.006	0.026	0.006	0.006	0.031	0.006	0.009	0.065	0.009	0.02	0.262	0.02	0.02	0.297	0.02	0.012	0.15	0.012
F3	0.004	0.004	0.017	0.006	0.006	0.025	0.006	0.006	0.03	0.009	0.009	0.064	0.02	0.02	0.263	0.02	0.019	0.298	0.012	0.012	0.151
G1	0.011	0.003	0.003	0.017	0.005	0.005	0.019	0.005	0.005	0.048	0.009	0.009	0.141	0.019	0.02	0.342	0.031	0.031	0.255	0.011	0.011
G2	0.003	0.011	0.003	0.005	0.018	0.005	0.005	0.019	0.005	0.009	0.049	0.009	0.02	0.141	0.02	0.031	0.34	0.031	0.011	0.254	0.011
G3	0.003	0.003	0.011	0.005	0.005	0.017	0.005	0.005	0.019	0.009	0.009	0.048	0.02	0.019	0.142	0.031	0.031	0.342	0.011	0.011	0.254

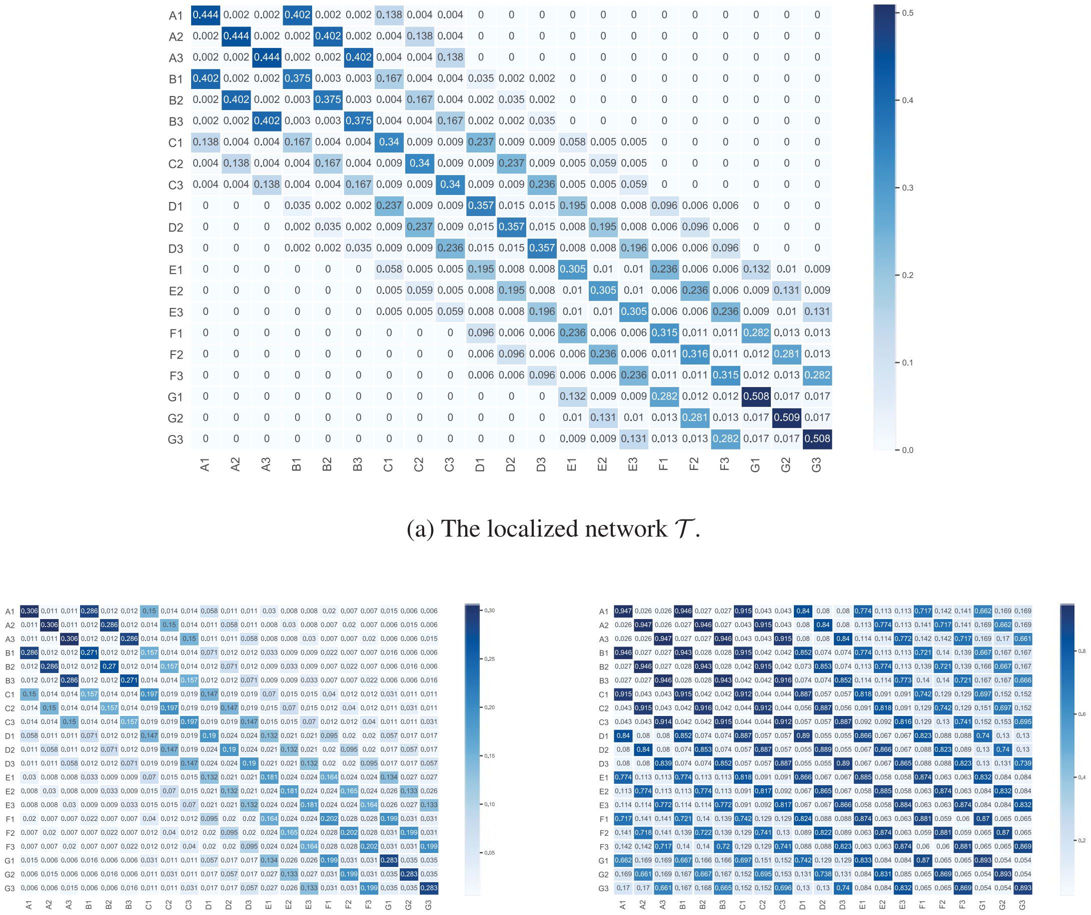
(a) Converged network $W_{t\to\infty}$ using $\alpha = 0.7$.



	A1	A2	A3	B1	B2	B3	C1	C2	C3	D1	D2	D3	E1	E2	E3	F1	F2	F3	G1	G2	G3
A1	0.958	0.021	0.021	0.956	0.022	0.022	0.936	0.032	0.032	0.902	0.049	0.049	0.848	0.076	0.076	0.81	0.094	0.095	0.771	0.114	0.114
A2	0.021	0.958	0.021	0.022	0.956	0.022	0.032	0.936	0.032	0.048	0.903	0.048	0.074	0.851	0.075	0.094	0.811	0.095	0.115	0.771	0.114
A3	0.021	0.021	0.958	0.022	0.022	0.956	0.032	0.032	0.936	0.049	0.049	0.902	0.074	0.074	0.852	0.092	0.092	0.815	0.115	0.115	0.77
B1	0.956	0.022	0.022	0.956	0.022	0.022	0.941	0.029	0.029	0.904	0.048	0.048	0.855	0.072	0.073	0.814	0.092	0.093	0.777	0.111	0.111
B2	0.022	0.956	0.022	0.022	0.956	0.022	0.029	0.941	0.029	0.048	0.905	0.048	0.071	0.857	0.072	0.092	0.815	0.093	0.112	0.776	0.111
B3	0.022	0.022	0.956	0.022	0.022	0.956	0.029	0.029	0.941	0.048	0.048	0.904	0.071	0.071	0.858	0.09	0.09	0.819	0.112	0.112	0.776
C1	0.936	0.032	0.032	0.941	0.029	0.029	0.95	0.025	0.025	0.93	0.035	0.035	0.877	0.061	0.062	0.846	0.076	0.077	0.806	0.097	0.097
C2	0.032	0.936	0.032	0.029	0.941	0.029	0.025	0.95	0.025	0.034	0.931	0.035	0.06	0.879	0.061	0.077	0.846	0.077	0.098	0.805	0.097
C3	0.032	0.032	0.937	0.029	0.029	0.941	0.025	0.025	0.95	0.035	0.035	0.93	0.06	0.06	0.88	0.075	0.075	0.85	0.098	0.097	0.804
D1	0.902	0.049	0.049	0.904	0.048	0.048	0.93	0.035	0.035	0.942	0.029	0.029	0.92	0.04	0.04	0.879	0.06	0.061	0.851	0.075	0.075
D2	0.049	0.902	0.049	0.048	0.904	0.048	0.035	0.93	0.035	0.029	0.942	0.029	0.04	0.921	0.04	0.06	0.879	0.061	0.076	0.849	0.075
D3	0.049	0.049	0.902	0.048	0.048	0.904	0.035	0.035	0.93	0.029	0.029	0.942	0.04	0.04	0.921	0.059	0.059	0.882	0.076	0.075	0.849
E1	0.849	0.076	0.076	0.855	0.072	0.073	0.877	0.062	0.061	0.919	0.04	0.04	0.935	0.033	0.033	0.926	0.037	0.037	0.889	0.056	0.056
E2	0.075	0.85	0.075	0.072	0.856	0.072	0.061	0.878	0.061	0.039	0.921	0.039	0.032	0.935	0.032	0.037	0.926	0.037	0.056	0.888	0.056
E3	0.075	0.075	0.85	0.072	0.072	0.856	0.061	0.062	0.877	0.04	0.04	0.92	0.032	0.032	0.936	0.037	0.037	0.927	0.056	0.056	0.888
F1	0.806	0.097	0.097	0.81	0.095	0.095	0.843	0.079	0.079	0.874	0.063	0.063	0.924	0.038	0.038	0.931	0.034	0.035	0.922	0.039	0.039
F2	0.096	0.807	0.097	0.094	0.811	0.095	0.078	0.844	0.078	0.061	0.878	0.061	0.038	0.924	0.038	0.035	0.931	0.035	0.039	0.921	0.039
F3	0.096	0.097	0.807	0.094	0.095	0.811	0.079	0.079	0.843	0.062	0.062	0.876	0.037	0.037	0.926	0.034	0.034	0.932	0.039	0.039	0.921
G1	0.754	0.123	0.123	0.76	0.12	0.12	0.79	0.105	0.105	0.835	0.083	0.082	0.88	0.06	0.06	0.916	0.042	0.042	0.944	0.028	0.028
G2	0.122	0.756	0.122	0.119	0.761	0.119	0.104	0.792	0.104	0.08	0.84	0.08	0.06	0.879	0.061	0.043	0.915	0.043	0.028	0.944	0.028
G3	0.122	0.122	0.757	0.119	0.119	0.762	0.105	0.105	0.791	0.081	0.081	0.838	0.059	0.058	0.883	0.042	0.042	0.916	0.028	0.028	0.944

(b) Category-based probability distributions for the test phase using $\beta_t = 4$.

Figure 2: The new network adjacency matrix when the regularization parameter is $\alpha = 0.7$ with the input matrix $W_0 = P$, which is given in Figure 1b. The test phase probabilities in Figure 2b are calculated by normalizing the weights for the specific category and then using the softmax function with parameter $\beta_t = 4$.



(a) The localized network $\mathcal{T}$.

(b) Converged network $W_{t\to\infty}$ using $\alpha = 0.7$.

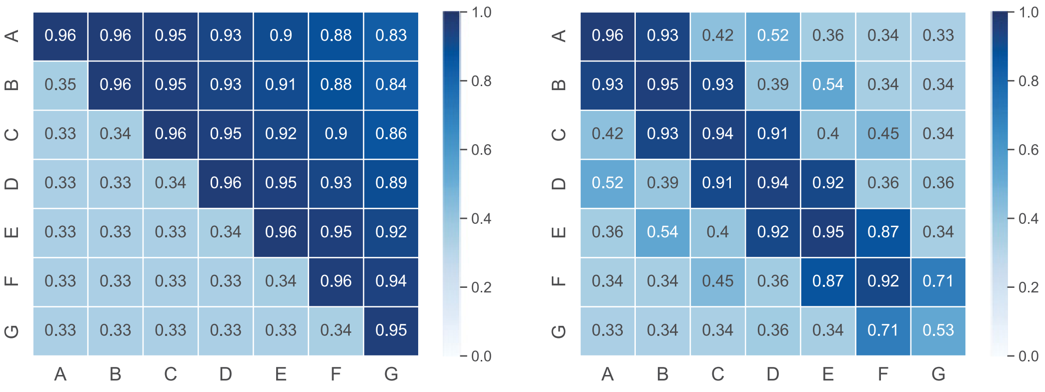
(c) Category-based probability distributions for the test phase using $\beta_t = 4$.

Figure 3: The new network adjacency matrix using an SNE update when the regularization parameter is $\alpha = 0.7$ and the input matrix $W_0 = P$, which is given in Figure 1b. The test phase probabilities in Figure 3c are calculated by normalizing the weights for the specific category and then using the softmax function with parameter $\beta_t = 4$.

4.2 Experiment 2: Isolating Symmetry and Transitivity. Two main differences between DNE and SNE are shown in this experiment. In this regard, we consider two extreme cases to isolate the symmetry and transitivity effects.

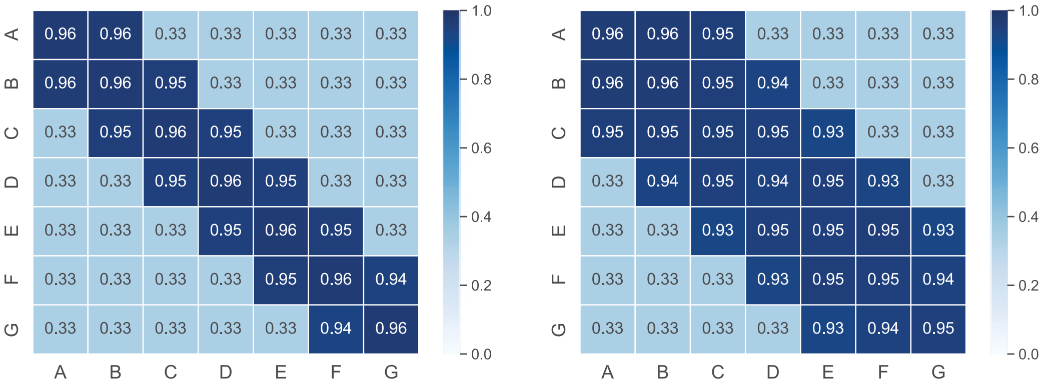
First, we isolate the effect of symmetry relations; in other words, we suppose that the agent is able to answer the transitive relations but unable to derive symmetry relations. For this, we set the parameters to $\gamma = 0.001$, $K = 0.01$, $\beta_h = 0.1$, $\beta_t = 4$, and $\alpha = 0.05$.

As illustrated in Figure 4a, the symmetry relations, and therefore the equivalence relations, can be altered by parameter K. However, in Figure 4b, due to the symmetric behavior of updates, the symmetry relations are exactly the same as the baseline relations, and the transitive and equivalence



(a) Final category-based results applying DNE (b) Final category-based results applying SNE

Figure 4: The probability of choosing correct pairs between categories when $\gamma = 0.001$, $K = 0.01$, $\beta_h = 0.1$, $\beta_t = 4$, and $\alpha = 0.05$. The reported values are calculated by taking the average over all relations in each category.



(a) Final category-based results applying DNE (b) Final category-based results applying SNE

Figure 5: The probability of choosing correct pairs between categories when $\gamma = 0.001$, $K = 1$, $\beta_h = 0.1$, $\beta_t = 4$, and $\alpha = 0$.

relations are altered by setting $K = 0.01$. We can conclude that a DNE-type agent can handle nonsymmetric relations, but the SNE agent is unable to control symmetry relations independently.

Next, we simulate a scenario in which the agent learns the baseline relations, but no transitive relation is derived. Suppose the symmetry relations are derived perfectly, so that we only isolate the transitive relations. Let the parameters of such an agent be $\gamma = 0.001, K = 1, \beta_h = 0.1, \beta_t = 4$, and $\alpha = 0$.

In Figure 5a, which uses the DNE method, the transitive and therefore equivalence relations are not formed, while the symmetry relations are strong. In Figure 5b, we see that the one-node relations such as AC and BD are derived in SNE. This is expected due to the definition of $\mathcal{T}$. In the EPS model, though, we are seeking to control all the transitive and equivalence relations.

Table 2: The Average of Required Repetition of Training Blocks until Reaching Mastery Criterion Ratio 0.9 When $\gamma = 0.001$, $K = 1$, $\beta_t = 4$, and $\alpha = 0.05$ for Three Values of $\beta_h = 0.2, 0.1,$ and 0.05.

	Number of Trials per Relation						Time		
Training	*AB*	*BC*	*CD*	*DE*	*EF*	*FG*	$\beta_h = 0.2$	$\beta_h = 0.1$	$\beta_h = 0.05$
AB	48						2.133	3.423	5.907
BC	24	24					2.885	4.757	8.751
CD	12	12	24				2.959	4.977	9.641
DE	9	9	9	24			2.791	4.661	9.469
EF	6	6	6	6	24		2.992	5.208	11.736
FG	3	3	3	6	9	24	3.008	5.339	12.978
Baseline maintenance	3	3	3	3	3	3	1.038	1.407	7.561

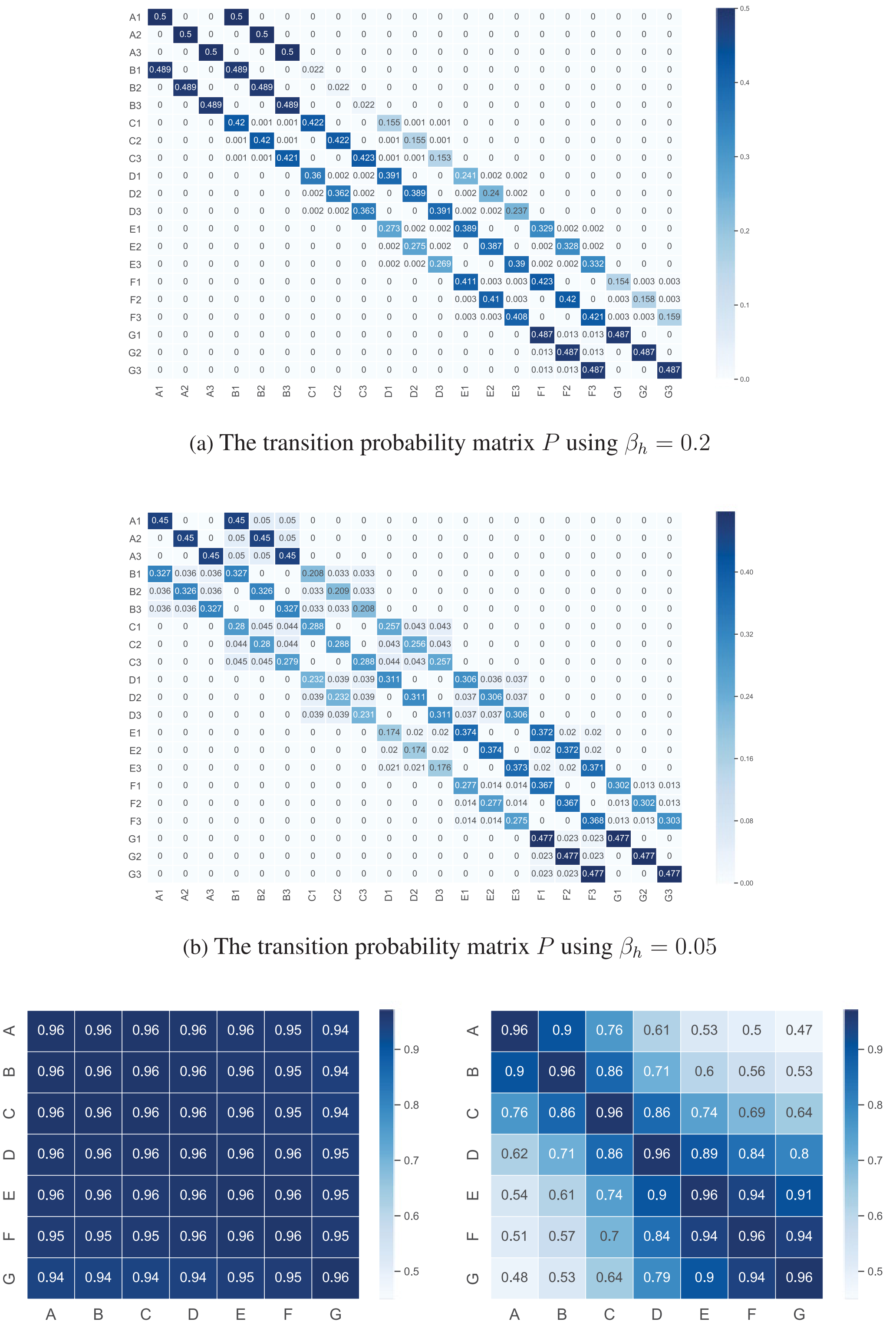
Therefore, since SNE is not an appropriate method for controlling symmetry and transitivity completely, we consider DNE as the main approach in this letter to cover more general cases, such as those with weak symmetry relations or weak transitivity relations. In the rest of the simulations, we report just the results for the DNE method.

4.3 Experiment 3: Effect of the β_h Parameter. The softmax function parameter β_h is used in the training phase for checking the mastery criterion as well as computing the transition matrix from W_h. As reported in Table 2, a higher value of β_h causes the agent to be able to pass the training phase faster, while for smaller values of β_h, it takes many more iterations to pass the training phase and learn baseline relations. Table 2 presents the learning speed for three values of $\beta_h = 0.2, 0.1,$ and 0.05 when $\gamma = 0.001$, $K = 1$, $\beta_t = 4$, and $\alpha = 0.05$.

Table 2 shows that parameter β_h can be used to control the learning speed. For instance, an agent with $\beta_h = 0.2$ learns *AB* relations by repeating the training blocks 2.1 times on average. This value will be 3.4 for $\beta_h = 0.1$ and 5.9 for $\beta_h = 0.05$.

Another effect of β_h appears in computing the probability matrix and, consequently, the final network shape. In Figure 6, we report the *P* matrix and the computed nodal effect in the test phase for two choices of $\beta_h = 0.2$ and $\beta_h = 0.05$ when we keep all parameters similar: $\gamma = 0.001$, $K = 1$, $\beta_t = 4$, and $\alpha = 0.05$.

By comparing Figures 6a and 6b, we notice that the probability of direct relations are weaker when $\beta_h = 0.05$. Since this matrix is considered as W_0, the input matrix to the NE iterative method, the final results will be altered. In Figure 6c, the nodal effect is negligible, and all the transitive and equivalence relations are formed equally well as baseline relations. Figure 6d, however, shows the nodal effect and the agent's weak performance in



(a) The transition probability matrix P using $\beta_h = 0.2$

(b) The transition probability matrix P using $\beta_h = 0.05$

(c) Final category-based probability of correct choice in the test phase when $\beta_h = 0.2$

(d) Final category-based probability of correct choice in the test phase when $\beta_h = 0.05$

Figure 6: Comparison of probability matrix out of training and final category-based probability of correct choice in the test phase for two choices of $\beta_h = 0.2$ and $\beta_h = 0.05$, when $\gamma = 0.001$, $K = 1$, $\beta_t = 4$, and $\alpha = 0.05$.

Table 3: The Average of Required Repetition of Training Blocks until Reaching Mastery Criterion Ratio 0.9 When $K = 1$, $\beta_h = 0.1$, $\beta_t = 4$, and $\alpha = 0.05$ for Three Values of $\gamma = 0, 0.001,$ and 0.005.

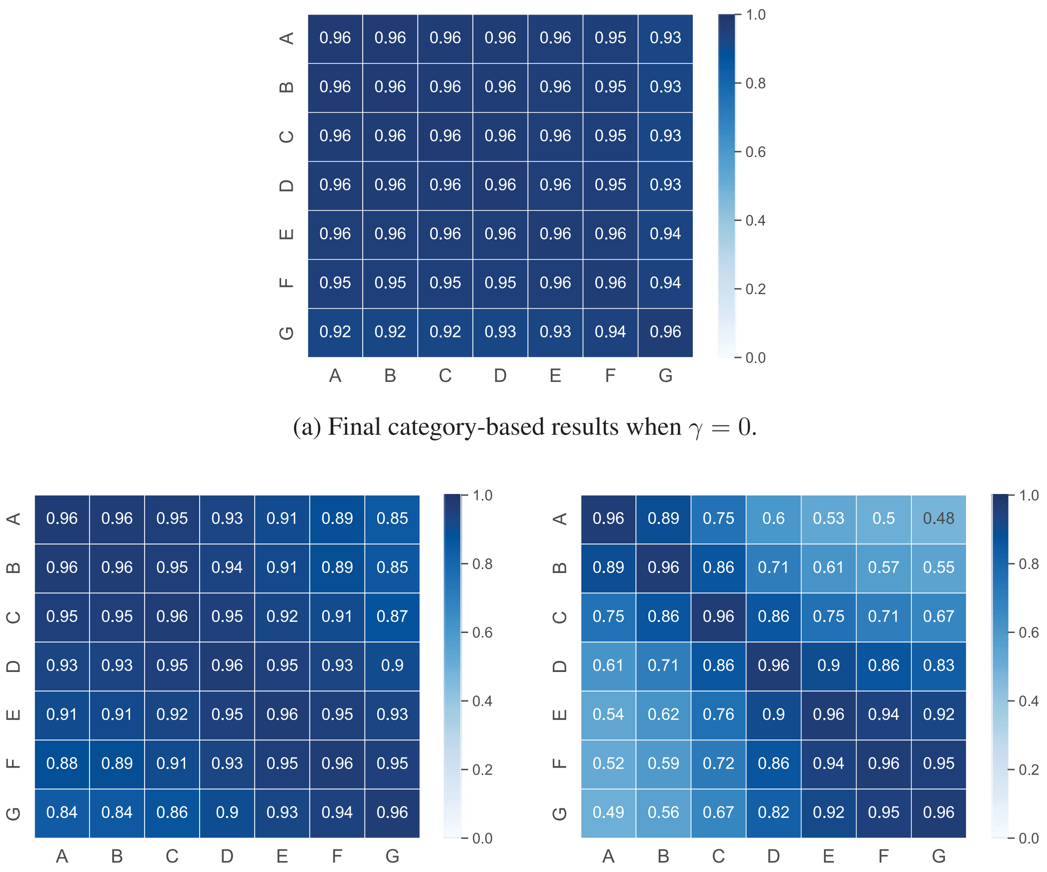
	Number of Trials per Relation						Time		
Training	*AB*	*BC*	*CD*	*DE*	*EF*	*FG*	$\gamma = 0.0$	$\gamma = 0.001$	$\gamma = 0.002$
AB	48						3.318	3.452	3.580
BC	24	24					4.391	4.703	5.088
CD	12	12	24				4.570	4.951	5.584
DE	9	9	9	24			4.200	4.654	5.514
EF	6	6	6	6	24		4.649	5.190	6.951
FG	3	3	3	6	9	24	4.637	5.324	7.884
Baseline maintenance	3	3	3	3	3	3	1.089	1.414	7.281

relations with a higher nodal number. We conclude that β_h can be used for controlling both the speed of learning and the nodal effect. In other words, if we fix all other parameters than β_h, the smaller value of β_h results in slower learning and a lower chance of forming transitive and equivalence relations with a higher nodal number. It is noteworthy that the effects of β_h and γ are somehow intertwined. As we see in experiment 4, γ also controls the learning speed and nodal effect. Indeed, if the agent does not forget at all, that is, $\gamma = 0$, then β_h controls just the speed of learning. However, $\gamma = 0$ is not a plausible choice for replication of human behavior.

4.4 Experiment 4: Effect of the γ Parameter. Mofrad et al. (2020) studied, the effect of γ in the training phase of EPS agents, where learning speed can be adjusted via γ. In Table 3, the average number of repetitions at each stage is provided for three choices: $\gamma = 0, 0.001,$ and 0.005. There is a general trend that increasing the forgetting factor will increase the repetition times in all stages. But the rates of increase for later stages and the baseline maintenance are different. The explanation is that the forgetting factor affects the initial learned relations more since at the final blocks, we have fewer of them. In other words, in the final blocks, we have fewer trials of them, and thus the forgetting factor will cause a stronger adverse impact. This is why we need around seven iterations of the maintenance phase when $\gamma = 0.002$, while we need just one iteration by removing the forgetting factor, $\gamma = 0$.

The forgetting factor will affect the final shape of h-values network W_h, and therefore for similar parameters, we have different probability matrices and final outcomes in the test phase. Figure 7 provides the final results of the testing phase for three different values of the forgetting factor: $\gamma = 0, 0.001, 0.002$.

When $\gamma = 0$ (see Figure 7a), there is no forgetting, and therefore the training order does not matter and all the relations are considered equally the



	A	B	C	D	E	F	G
A	0.96	0.96	0.96	0.96	0.96	0.95	0.93
B	0.96	0.96	0.96	0.96	0.96	0.95	0.93
C	0.96	0.96	0.96	0.96	0.96	0.95	0.93
D	0.96	0.96	0.96	0.96	0.96	0.95	0.93
E	0.96	0.96	0.96	0.96	0.96	0.96	0.94
F	0.95	0.95	0.95	0.95	0.96	0.96	0.94
G	0.92	0.92	0.92	0.93	0.93	0.94	0.96

(a) Final category-based results when $\gamma = 0$.

	A	B	C	D	E	F	G
A	0.96	0.96	0.95	0.93	0.91	0.89	0.85
B	0.96	0.96	0.95	0.94	0.91	0.89	0.85
C	0.95	0.95	0.96	0.95	0.92	0.91	0.87
D	0.93	0.93	0.95	0.96	0.95	0.93	0.9
E	0.91	0.91	0.92	0.95	0.96	0.95	0.93
F	0.88	0.89	0.91	0.93	0.95	0.96	0.95
G	0.84	0.84	0.86	0.9	0.93	0.94	0.96

(b) Final category-based results when $\gamma = 0.001$

	A	B	C	D	E	F	G
A	0.96	0.89	0.75	0.6	0.53	0.5	0.48
B	0.89	0.96	0.86	0.71	0.61	0.57	0.55
C	0.75	0.86	0.96	0.86	0.75	0.71	0.67
D	0.61	0.71	0.86	0.96	0.9	0.86	0.83
E	0.54	0.62	0.76	0.9	0.96	0.94	0.92
F	0.52	0.59	0.72	0.86	0.94	0.96	0.95
G	0.49	0.56	0.67	0.82	0.92	0.95	0.96

(c) Final category-based results when $\gamma = 0.002$

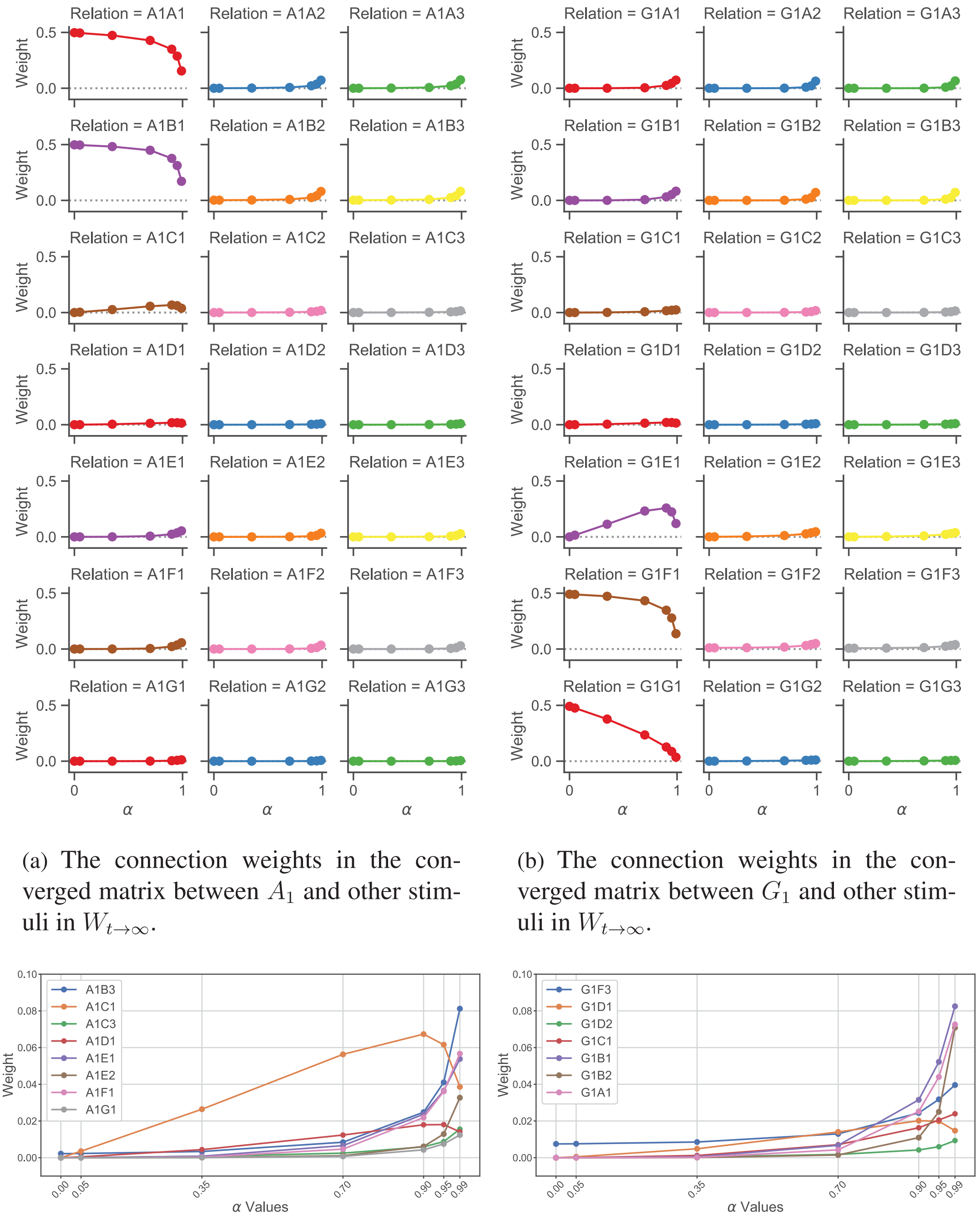
Figure 7: Probability of choosing correct pairs between categories when $K = 1$, $\beta_h = 0.1$, $\beta_t = 4$, and $\alpha = 0.05$ for three forgetting factor values: $\gamma = 0, 0.001$, and 0.002.

same. In Figure 7b, all the relations are formed but we can easily notice the nodal effect. For instance, if we test the *AB* relation, the probability of a correct choice by the agent is 0.96, while it is about 0.85 for *AG* with five nodes in between. Figure 7c shows that a higher forgetting factor can be used to model impaired equivalence class formation. If we test the agent with the *AB* relation, the probability of a correct choice would be 0.89, while it is about 0.48 for *AG*. Comparing the correct choice probabilities for *AB* and *FG* (0.89 for *AB* versus 0.95 for *FG*) shows the importance of training order in this setting. The agent forgets the initial stage relations, and these relations need to be repeated. If the training trial blocks are totally separate, as in experiment 1 in Mofrad et al. (2020), the initial trained relations drop dramatically with a high forgetting factor.

To show the importance of testing order in the model, similar to the SE literature, we simulate the testing phase with different test orders so the trials that appear late in the testing phase have weaker results when the forgetting factor is high. Here, for simplicity, we calculate the probability distribution for different test trials and evaluate the agent behavior based on them. This means the forgetting factor is not effective on the test results in the current simulations. However, the forgetting factor can be used by defining β_t as a function of time and γ to model the forgetting in the testing phase of E-EPS. Another argument is that the forgetting might affect the network; in this case, the network weights must be updated in a way to keep each row summing to one. Therefore, it is not as straightforward as the EPS where the matrix with h-values is the basis for the testing phase.

4.5 Experiment 5: Effect of the α Parameter. This parameter shapes the final representation of the clip network (see appendix A for a theoretical discussion). A smaller value of α biases the converged matrix $W_{t\rightarrow\infty}$ to keep the connections from W_0 stronger, while a larger value of α enhances transitive relations. In the case of $\alpha = 0$, as represented in Figure 5a, there is no enhancement in the network using DNE. Figures 8a and 8b, respectively, represent the connection values from A_1 and G_1 to other stimuli in the converged network for $\alpha = 0, 0.05, 0.35, 0.7, 0.9, 0.95,$ and 0.99, when $\gamma = 0.001$, $K = 1$, and $\beta_h = 0.1$.

As depicted in Figure 8, smaller values of α keep the relations in the input network (i.e., trained relations together with symmetry and reflexivity) stronger. On the other hand, a higher α value reinforces the transitive and equivalence relations. For each α value, the connection weights for all relations must sum to one; for instance, the values for $\alpha = 0.9$ in all subplots of Figure 8a sum to one as they show the transition probability from A_1 to all other points when using $\alpha = 0.9$. As a result, increasing the values for transitive relations means a decrease in initial relations (see the decrease in A_1A_1, A_1B_1 relation weights and the increase in other values, say, A_1C_1 and A_1G_1). Along with construction and enhancing the desired relations (see the first columns in Figures 8a and 8b), the undesired relations are also constructed and enhanced to some extent. This can be explained by the fact that the values for undesired relations such as A_1B_2, A_1B_3, G_1F_2, and G_1F_3 are not zero in the initial matrix since the training criterion was set to 0.9. These values could enhance undesired relations, especially when α is higher. For instance, as depicted in Figure 8c, the connection weight for the A_1C_1 relation, which is a desired relation, decreases for α values higher than 0.9. Similarly, the connection weight for the A_1D_1 relation decreases at $\alpha = 0.99$ in comparison with $\alpha = 0.9, 0.95$. The connection weight for the A_1B_3 relation, which has a very small weight in the beginning (i.e., when $\alpha = 0$), increases with α with acceleration in the rate of change for α values greater than 0.7. A_1C_3 and A_1E_2 are two sample relations that are undesirable and get enhanced during the diffusion process as a function of α value. The same kind of behavior can be observed for relations from G_1. In Figure 8d, the relation G_1D_1 increases as desired, but when α is too high ($\alpha = 0.95, 0.99$), it starts to decrease. Undesired relations such as G_1F_3 and G_1D_2 are enhanced with



(a) The connection weights in the converged matrix between A_1 and other stimuli in $W_{t\to\infty}$.

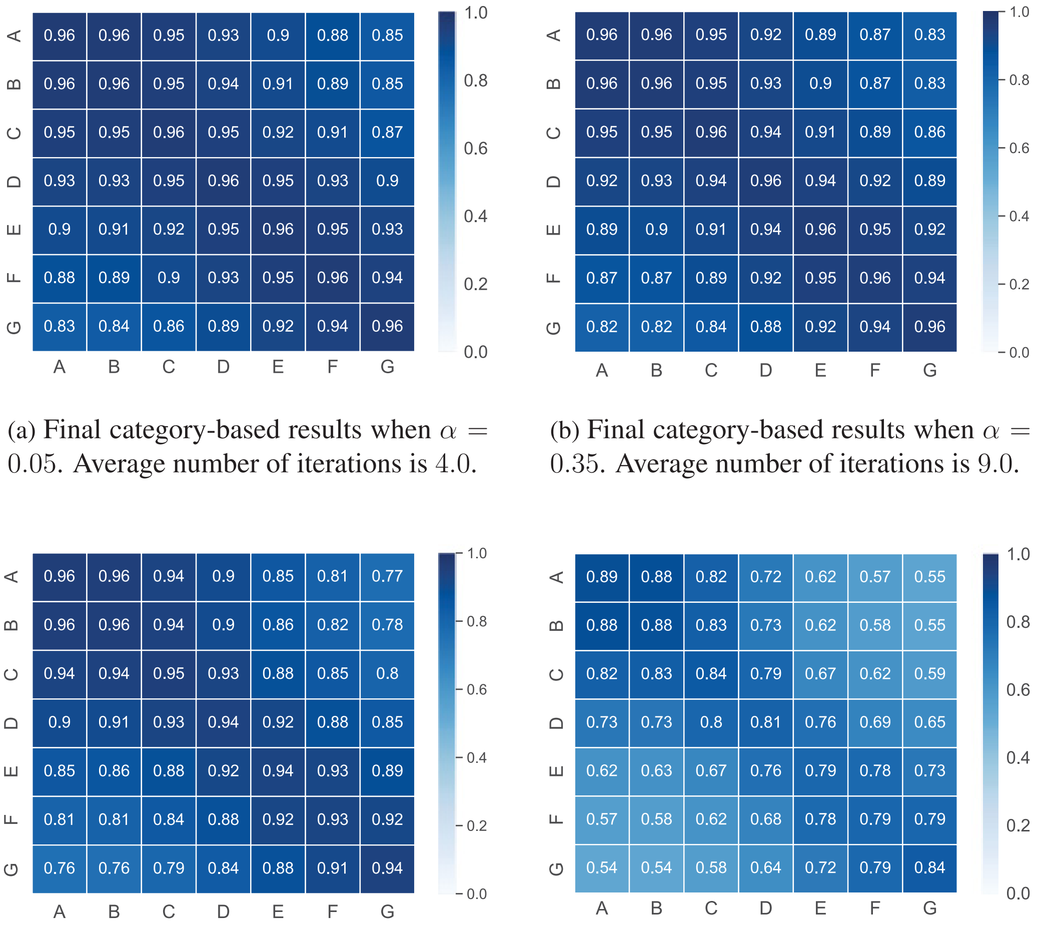
(b) The connection weights in the converged matrix between G_1 and other stimuli in $W_{t\to\infty}$.

(c) A comparison between behavior of some desired and undesired relations between A_1 and other stimuli based on different α values.

(d) A comparison between behavior of some desired and undesired relations between G_1 and other stimuli based on different α values.

Figure 8: The connection weights in the converged matrix $W_{t\to\infty}$ for A_1 and G_1 for $\alpha = 0, 0.05, 0.35, 0.7, 0.9, 0.95, 0.99$, when $\gamma = 0.001$, $K = 1$, and $\beta_h = 0.1$.

a higher rate when α approaches one. Therefore, an inappropriate choice of α could be destructive; in this example, a higher value of α than 0.9 sounds inappropriate.



(a) Final category-based results when $\alpha = 0.05$. Average number of iterations is 4.0.

(b) Final category-based results when $\alpha = 0.35$. Average number of iterations is 9.0.

(c) Final category-based results when $\alpha = 0.7$. Average number of iterations is 23.96.

(d) Final category-based results when $\alpha = 0.95$. Average number of iterations is 102.0.

Figure 9: Probability of choosing correct pairs between categories when $\gamma = 0.001$, $K = 1$, $\beta_h = 0.1$, and $\beta_t = 4$ for $\alpha = 0.05, 0.35, 0.7,$ and 0.95.

Different α values and therefore different configurations of the $W_{t\to\infty}$ matrix result in different testing performance. In Figure 9, we report the testing results for $\alpha = 0.05, 0.35, 0.7, 0.95$ when $\gamma = 0.001$, $K = 1$, $\beta_h = 0.1$, and $\beta_t = 4$.

We observe that the probabilities of choosing correct relations in Figures 9c and 9d, respectively, for $\alpha = 0.05$ and $\alpha = 0.35$ are almost the same. In Figure 9a, when $\alpha = 0.7$, the transitive and equivalence relations are affected negatively. In Figure 9d, we see from the converged transition matrix that values for all the relations have decreased. Moreover, for smaller values of α, the convergence of the network needs fewer iterations; compare 4, 9, 23, and 102 for, respectively, $\alpha = 0.05, 0.35, 0.7$, and 0.95. For more details in α parameter effect, see Table 4, where the connection weights of AB and AG in $W_{t\to\infty}$ for different α choices, along with the calculated probabilities based on three choices of $\beta_t = 1, 4, 8$, are reported.

Table 4: The Simultaneous Effect of α and β_t Values on the Test Results for AB and AG Relations.

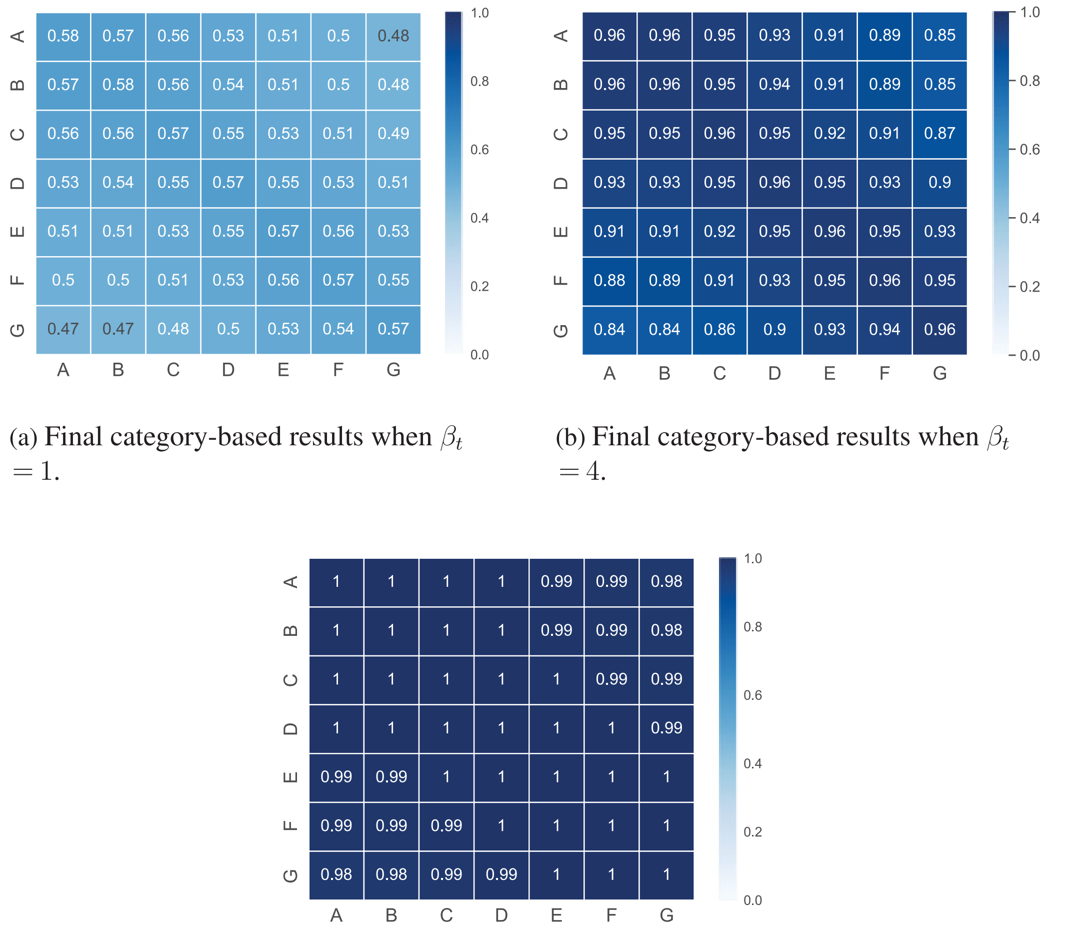
		Baseline Relation AB			Derived Relation AG		
(α, β_t)		A_1B_1	A_1B_2	A_1B_3	A_1G_1	A_1G_2	A_1G_3
$\alpha = 0$	$W_{t\to\infty}$	0.49837	0.00163	0.00163	0	0	0
	$W_{t\to\infty_C}$	0.99350	0.00325	0.00325	0	0	0
	$\beta_t = 1$	0.57134	0.21419	0.21447	0.33333	0.33333	0.33333
	$\beta_t = 4$	0.9619	0.01904	0.01906	0.33333	0.33333	0.33333
	$\beta_t = 8$	0.99925	0.00037	0.00037	0.33333	0.33333	0.33333
$\alpha = 0.05$	$W_{t\to\infty}$	0.49686	0.0017	0.0017	$4.1276e^{-08}$	$5.6875e^{-09}$	$8.6627e^{-09}$
	$W_{t\to\infty_C}$	0.9932	0.0034	0.0034	0.74202	0.10225	0.15573
	$\beta_t = 1$	0.57115	0.21429	0.21456	0.48349	0.25909	0.25743
	$\beta_t = 4$	0.96178	0.01909	0.01912	0.83865	0.08146	0.07989
	$\beta_t = 8$	0.99925	0.00037	0.00037	0.99049	0.00509	0.00442
$\alpha = 0.9$	$W_{t\to\infty}$	0.39782	0.02119	0.0223	0.003	0.00092	0.00112
	$W_{t\to\infty_C}$	0.90145	0.04802	0.05053	0.59524	0.18254	0.22222
	$\beta_t = 1$	0.51983	0.24069	0.23948	0.4146	0.29794	0.28746
	$\beta_t = 4$	0.91757	0.04108	0.04135	0.69558	0.17058	0.13384
	$\beta_t = 8$	0.99726	0.00137	0.00136	0.96297	0.02154	0.01549
$\alpha = 0.95$	$W_{t\to\infty}$	0.34433	0.03844	0.04031	0.00464	0.00185	0.00212
	$W_{t\to\infty_C}$	0.81387	0.090858	0.095278	0.53891	0.21487	0.24623
	$\beta_h = 1$	0.47784	0.2627	0.25945	0.39334	0.31058	0.29608
	$\beta_h = 4$	0.85289	0.0733	0.0738	0.61673	0.22268	0.16059
	$\beta_h = 8$	0.99183	0.0041	0.00407	0.92776	0.04397	0.02827

Notes: The $W_{t\to\infty}$ row reports the weights in the converged network. $W_{t\to\infty_C}$ refers to the input weights conditioned based on the category that softmax function uses to generate the probability distribution. The C in the index of $W_{t\to\infty_C}$ refers to the conditional weights for the category calculated with Bayes' rule.

4.6 Experiment 6: Effect of the β_t Parameter. To study the effect of β_t, first we keep other parameters fixed ($\gamma = 0.001$, $K = 1$, $\beta_h = 0.1$, $\alpha = 0.05$) and simulate the agent behavior for $\beta_t = 1, 4, 8$ (see Figure 10).

We see a decrease in all types of relations by decreasing the value of β_t. In Figure 10a, when $\beta_t = 1$, all relations, including baseline relations, become weaker. When $\beta_t = 4$ in Figure 10b, we see that the relations are well formed across all nodal numbers. Figure 10c shows that with a higher value of $\beta_t = 8$, all the relations are almost completely formed. This experiment illustrates that by changing β_t, one can control the agent performance in the testing phase and even impair the baseline relations. In Table 4, we take a closer look at the simultaneous effect of α and β_t when $\gamma = 0.001$, $K = 1$, $\beta_h = 0.1$.

In Table 4, baseline relation AB and transitive relation AG with nodal number five are addressed. We use the conditioned weights (row $W_{t\to\infty_C}$) as the input vector to the softmax function to generate the probability distribution for the test phase. When $\alpha = 0$, there is no NE, and any choice of β_t results in an equal probability of all relations in AG. However, β_t



(a) Final category-based results when $\beta_t = 1$.

(b) Final category-based results when $\beta_t = 4$.

(c) Final category-based results when $\beta_t = 8$.

Figure 10: Probability of choosing correct relations between categories when $\gamma = 0.001$, $K = 1$, $\beta_h = 0.1$, and $\alpha = 0.05$ for $\beta_t = 1, 4, 8$.

could affect the AB relation so that the performance of the agent is very poor (it chooses A_1B_1 with probability 0.57134 for $\beta_t = 1$) or very strong (it chooses A_1B_1 with probability 0.99925 for $\beta_t = 8$). When $\alpha = 0.05$, $W_{t\to\infty}$ is achieved after about just four iterations. We observe an insignificant reduction in the A_1B_1 weight in $W_{t\to\infty}$ (from 0.49837 to 0.49686) and an insignificant increase in A_1B_2, A_1B_3, A_1G_1, A_1G_2, and A_1G_3. Interestingly, since we use conditioned weights and apply a softmax function, very tiny values for AG in $W_{t\to\infty}$ transfer into noticeable values when conditioned, which could show the formation of derived relations. For instance, with $\beta_t = 4$, $(A_1G_1, A_1G_2, A_1G_3)_{W_{t\to\infty}} = (4.1276e^{-08}, 5.6875e^{-09}, 8.6627e^{-09})$ is transformed to (0.74202, 0.10225, 0.15573) and when softmax is used, it is converted into (0.83865, 0.08146, 0.07989), that is, an A_1G_1 relation is formed for the agent. This means a small value of α and, consequently, a few steps of NE could produce the desired network with an appropriate choice of β_t.

Table 5: The Training Order for OTM.

	Number of Trials per Relation					
Training	*AB*	*AC*	*AD*	*AE*	*AF*	*AG*
AB	48					
AC	24	24				
AD	12	12	24			
AE	9	9	9	24		
AF	6	6	6	6	24	
AG	3	3	3	6	9	24
Baseline maintenance	3	3	3	3	3	3

If we consider higher values of α, we see that the weight of baseline relation A_1B_1 in $W_{t\to\infty}$ is reduced, but all other relations are enhanced.

It is also noteworthy that increasing the value of A_1G_1, which happens with a higher choice of α, is not equivalent to better performance in the testing phase as reported in Table 4.

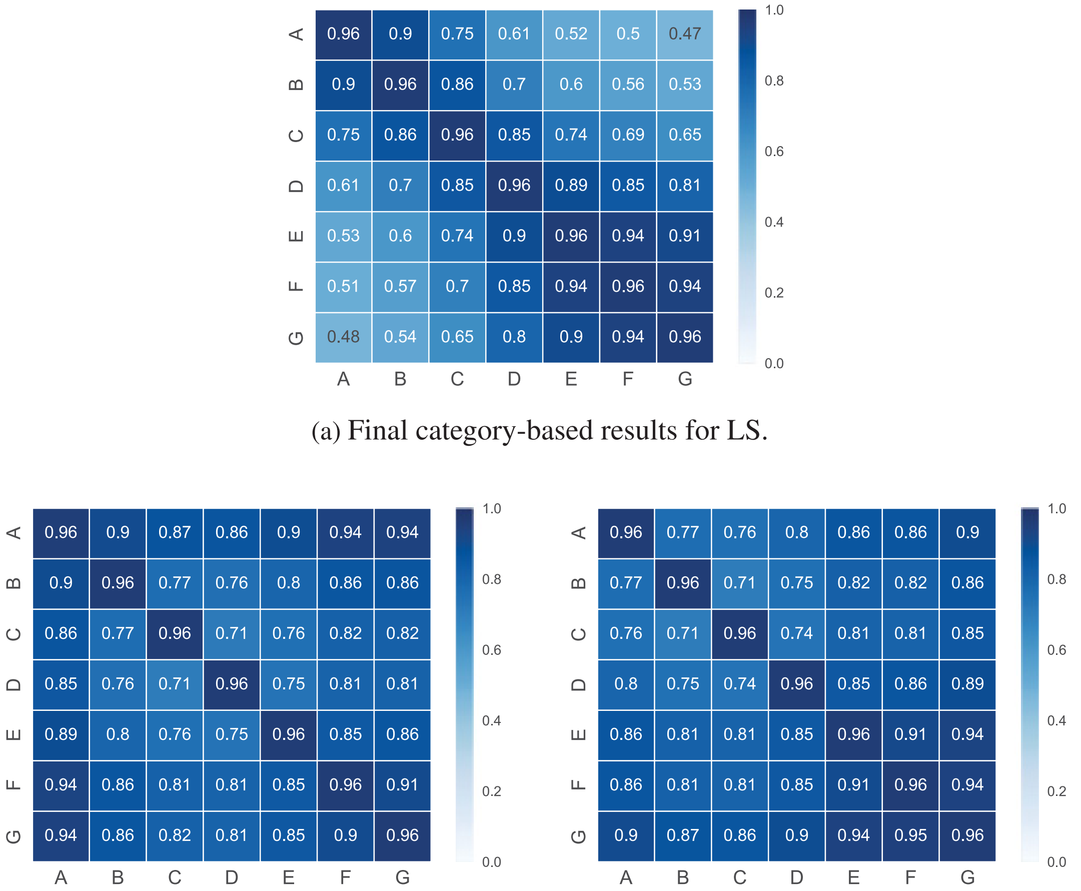
The reason is that NE changes the proportion of weights in $W_{t\to\infty}$, which affects the conditioned vector in favor of undesired options (see the $W_{t\to\infty_C}$ values), and, finally, the probability of a correct choice computed through the softmax function is reduced. For instance, when $\alpha = 0.05$, the A_1G_1 weight is $4.1276e^{-08}$, but its proportion in the conditioned vector is 0.74202. For $\alpha = 0.95$, the A_1G_1 weight is 0.00464, which is much higher than $\alpha = 0.05$, but its proportion in the conditioned vector is 0.53891, which is less than the case with $\alpha = 0.05$. So different configurations of α and β_t could produce different behaviors on request.

4.7 Experiment 7: Studying the Training Order: Comparing LS, MTO, and OTM. There are many studies on the differences between LS, OTM, and MTO training structures (see, e.g., Arntzen et al., 2010; Arntzen & Hansen, 2011; Arntzen, 2012). In this experiment, we rearrange the training blocks from LS in Table 1 to similar training stages for OTM and MTO training structures, represented in Tables 5 and 6, respectively. For the OTM training structure, the training relations in order are AB, AC, AD, AE, AF, and AG. For the MTO training structure, the training relations in order are AG, BG, CG, DG, EG, and FG.

The LS, OTM, and MTO training structures can be studied in various levels and with several parameter assemblies. But the aim of this experiment is to show the potential of the proposed E-EPS model in reflecting the differences between the LS, OTM, and MTO training structures reported in the literature. Figure 11 reports the results of the final testing phase of the three cases for an agent with parameters $\gamma = 0.001$, $K = 1$, $\beta_h = 0.05$, $\alpha = 0.05$, and $\beta_t = 4$.

Table 6: The Training Order for MTO.

	Number of Trials per Relation					
Training	*AG*	*BG*	*CG*	*DG*	*EG*	*FG*
AG	48					
BG	24	24				
CG	12	12	24			
DG	9	9	9	24		
EG	6	6	6	6	24	
FG	3	3	3	6	9	24
Baseline maintenance	3	3	3	3	3	3



(a) Final category-based results for LS.

(b) Final category-based results for OTM.

(c) Final category-based results for MTO.

Figure 11: Probability of choosing the correct relations between categories when $\gamma = 0.001$, $K = 1$, $\beta_h = 0.05$, $\alpha = 0.05$, and $\beta_t = 4$ for LS, MTO, and OTM.

According to Figure 11a, the agent performance when the LS is used is not satisfactory for higher nodal numbers. The weakest value, 0.47, belongs to *AG*. The equivalence classes are not formed in this case. Figure 11b shows better performance where the weakest connections are for *CD* and *DC* and equal 0.71. This minimum value is also found in Figure 11c but for relations *BC* and *CB*. So in this experiment, the overall results in terms of formation of equivalence classes are the same for MTO and OTM, but due to the order of training, the agent might exhibit different performance for specific relations in MTO and OTM training structures. For instance, the calculated probability for an *FA* relation in OTM is 0.94, and in MTO it is 0.86. Calculated probability for the *DE* relation in OTM is 0.75, while in MTO it is 0.85.

It is noteworthy that the training times—that is, the numbers of repetitions of each block before mastery in all three cases for all training procedures—are similar. This can be explained by the independence of designing baseline relations. The reported results in Figure 11 confirm that our model shows better performance in the OTM and MTO cases in comparison with LS (Arntzen et al., 2010; Arntzen & Hansen, 2011; Arntzen, 2012).

5 Conclusion

The main contribution of this letter is to offer a new perspective in the formation of SE classes in a recently introduced model, EPS. EPS is a modified version of the PS model (Briegel & De las Cuevas, 2012) and can be seen as an RL agent that has a directed, weighted network of clips. Each clip represents a remembered stimulus that is added to the clip network during the training phase.

To replicate the test phase of SE by examining the agent's ability to encounter new relations that can be derived from baseline relations, the EPS model relies on some type of likelihood reasoning whenever tested via an MTS trial. In other words, in the EPS model, derived relations were calculated on demand in the testing phase trials, but the new approach to the testing phase is offline and relies on memory retrieval during the testing phase rather than on complex logical processing. Derived relations in the new model, E-EPS, are achieved by applying an iterative diffusion process, network enhancement (NE; Wang et al., 2018). During the NE phase, the structure of the clip network changes where indirect relations get enhanced. The NE is a denoising method, and one way to interpret the model is to consider a typical memory as a less noisy memory, while a disabled memory is a noisy memory that cannot form equivalence relations. Since in the NE, connections are bidirectional, we refer to it is as symmetric network enhancement (SNE) in this letter. We further modify the SNE and propose directed network enhancement (DNE) in which the connections are directed and where we can control the agent's ability to derive transitivity and symmetry. One might use SNE in studying SE formation with the assumption that all the relations are bidirectional and transitive and equivalence relations are formed. DNE is a better option to replicate real experiments with the possibility of nonformation of classes and nonsymmetric relations.

In the simulation part, we study the role of parameters on agent performance and show that the model is able to replicate either a typical memory or a disabled memory with different learning and forgetting rates and accomplish the trial tasks in the testing phase. We also compare the main training structures, LS, MTO, and OTM, and notice the better outcome of MTO and OTM training structures than that of LS, which is consistent with evidence from the behavioral analysis literature. Many other configurations can be considered in simulations. For instance, we consider $K = 1$ to reduce the variety of results, and to study each parameter, we fixed all the other parameters.

Another alternative is to execute the NE phase during training rather than merely at the end of the training. The argument would be that brain does not wait until the end of training to start the process of formation of these relations. Although this might sound a like plausible argument and can be easily added to the model, we avoid NE during training. The most obvious reason is to keep the model simple, with fewer computations. Because we are studying agent behavior, the timing of events inside the brain is not our priority. Moreover, baseline relations are independent and not derived from each other, so there is no need to update them earlier when the formation of relations is tested in the testing phase. However, as discussed in section 3.1, these updates could be analogous to the replay in the brain that generates a predictive map in an offline process.

The probability distribution over comparison stimuli in the test trial is calculated based on the direct links in the updated clip network. It is similar to the EPS in the sense that whenever there are links between the sample stimulus and comparison stimuli, the probabilities are calculated based on the h-values by averaging or using a softmax function. In E-EPS, however, there are links through the entire network updated by the NE process, and therefore no extra calculation is made. Although one might still consider the random walk on the network similar to the PS model, the cyclic nature of the network in E-EPS might generate problems, and extra conditions (such as gating) might be necessary. We avoid this scenario, since the calculated weights are based on the random walk and diffusion, and we consider these cached links at the decision time. The EPS and E-EPS could be developed further to model more complex tasks with more sophisticated structures as the PS model offers. For instance, we might use compound stimuli and benefit from a PS model with associative learning (Briegel & De las Cuevas, 2012), or a multilayer memory clip where the agent is able to generate and add wildcard to the memory (Melnikov et al., 2017). Such multi-layer PS agent has been further developed to address abstract compositional concepts which is closer to the concept of SE (Ried, Eva, Müller, & Briegel, 2019). The mathematical understanding of the properties of the converged network that guarantees the converged solution is an advantage of NE over other network denoising methods. DNE maintains many properties of SNE with the advantage of controlling the formation of symmetry and transitivity in the E-EPS model. Finally, it is worth mentioning that we choose NE as the source of inspiration for updating the network clip, since in the updates, there is no requirement for supervision or prior knowledge. After the training phase, we have a clip network without further feedback or supervision. Hence, NE provides a proper solution with an emphasis on the indirect paths, which is what we have in derived relations.

Appendix A: Theoretical Analysis of Directed Network Enhancement

In this appendix, we explain why the proposed diffusion process in equation 3.1 improves the results and can be used to form equivalence classes. Our theoretical analysis is mostly based on supplementary note 3 of Wang et al. (2018). However, since W_t in the DNE is not a symmetric doubly stochastic matrix, the proofs and discussions need to be revised for DNE. It is noteworthy that the largest eigenvalue of each right stochastic matrix, such as P, is 1, associated with eigenvector $\mathbf{1}$. We first prove that W_t remains right stochastic in each iteration of DNE and converges to a nontrivial equilibrium matrix. Then we show that DNE preserves the eigenvectors of the stochastic matrix W_0, but increases the gap between large eigenvalues and reduces the gap between small eigenvalues (see Figure 13). The larger eigengap in the final converged matrix $W_{t\to\infty}$, is associated with better equivalence class formation.

A.1 The Convergence of the DNE Process.

We show that W_t remains stochastic during the updates. By definition $W_0\mathbf{1} = \mathbf{1}$, for all-one eigenvector $\mathbf{1}$ associated with eigenvalue 1. We assume that $W_{t-1}\mathbf{1} = \mathbf{1}$ and show that the rows of W_t remain normalized:

$$\begin{aligned} W_t\mathbf{1} &= \alpha P W_{t-1} P\mathbf{1} + (1-\alpha)P\mathbf{1} \\ &= \alpha P W_{t-1}\mathbf{1} + (1-\alpha)P\mathbf{1} \\ &= \alpha P\mathbf{1} + (1-\alpha)P\mathbf{1} \\ &= P\mathbf{1} \\ &= \mathbf{1}. \end{aligned} \tag{A.1}$$

Now we show that W_t converges to a nontrivial equilibrium graph. A closed-form solution for the final, converged network can be achieved through induction. Consider the following expression for the network at iteration t:

$$W_t = \alpha^t P^t W_0 P^t + (1-\alpha)P\sum_{k=0}^{t-1}(\alpha P^2)^k. \tag{A.2}$$

This formula is similar to equation 6 of supplementary note 3 by Wang et al. (2018) where $\mathcal{T}$ is replaced by P and can be guessed by iterating the process for the first few steps:

1. Define $W_0 = W_{t=0}$. For $t = 1$, equation A.2 holds true:

$$W_{t=1} = \alpha P W_0 P + (1-\alpha)P$$

2. We assume equation A.2 holds true for iteration t. Then:

$$\begin{aligned} W_{t+1} &= \alpha P W_t P + (1-\alpha)P \\ &= \alpha P \left(\alpha^t P^t W_0 P^t + (1-\alpha)P \sum_{k=0}^{t-1} (\alpha P^2)^k \right) P + (1-\alpha)P \\ &= \alpha^{t+1} P^{t+1} W_0 P^{t+1} + (1-\alpha)P \sum_{k=0}^{t-1} (\alpha P^2)^{k+1} + (1-\alpha)P \\ &= \alpha^{t+1} P^{t+1} W_0 P^{t+1} + (1-\alpha)P \sum_{k=0}^{t} (\alpha P^2)^k, \end{aligned}$$

which satisfies equation A.2. Using geometric series when $t \to \infty$, we have this nontrivial equilibrium matrix:

$$W_{t\to\infty} = (1-\alpha)P(\mathcal{I} - \alpha P^2)^{-1}. \tag{A.3}$$

A.2 Spectral Analysis of DNE. We show that the DNE process does not change eigenvectors of the input matrix $W_0 = P$ but maps eigenvalues through a nonlinear function.

Suppose (λ, v) is the eigenpair of P. We know that the absolute value of eigenvalues of any stochastic matrix satisies the $|\lambda| \leq 1$ relation. Let the eigendecomposition of P be VDV^{-1}, where D is a diagonal matrix formed from eigenvalues of P and the columns of V are the corresponding eigenvectors of P. We have

$$\begin{aligned} W_{t\to\infty} &= (1-\alpha)P(\mathcal{I} - \alpha P^2)^{-1} \\ &= (1-\alpha)VDV^{-1}(\mathcal{I} - \alpha VDV^{-1}VDV^{-1})^{-1} \\ &= (1-\alpha)VDV^{-1}(VV^{-1} - \alpha VDV^{-1}VDV^{-1})^{-1} \\ &= (1-\alpha)VDV^{-1}\left(V(\mathcal{I} - \alpha D^2)V^{-1}\right)^{-1} \\ &= (1-\alpha)VDV^{-1}\left(V(\mathcal{I} - \alpha D^2)^{-1}V^{-1}\right) \\ &= (1-\alpha)V\left(D(\mathcal{I} - \alpha D^2)^{-1}\right)V^{-1} \\ &= V\left((1-\alpha)(D(\mathcal{I} - \alpha D^2)^{-1}\right)V^{-1} \\ &= VD'V^{-1}. \end{aligned}$$

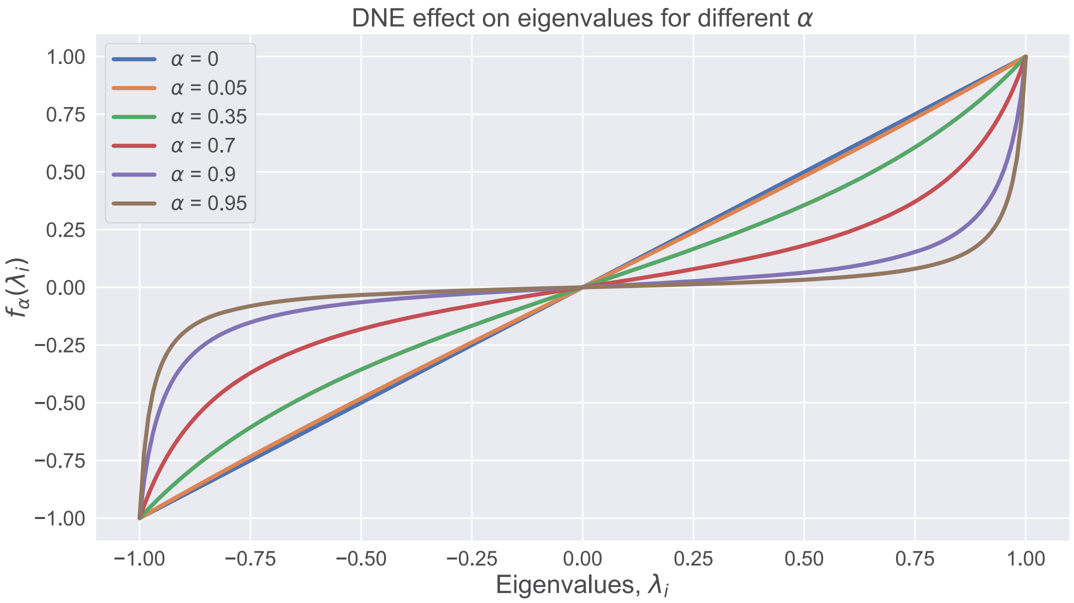


Figure 12: Role of α on the nonlinear transformation of eigenvalues using $f_\alpha(\lambda)$ in the DNE process.

This testifies that the DNE process keeps the eigenvectors unchanged, but the eigenvalues become $D'_{ii} = \frac{(1-\alpha)\lambda_i}{1-\alpha\lambda_i^2}$. Therefore, the DNE process functions nonlinearly on the eigenvalues of the input matrix, that is, the final converged matrix, $W_{t\to\infty}$, transforms (λ, v) to $(f_\alpha(\lambda), v)$, where $f_\alpha(\lambda) = \frac{(1-\alpha)\lambda}{1-\alpha\lambda^2}$. It is trivial that $f_\alpha(\lambda)(0) = 0$, $f_\alpha(\lambda)(1) = 1$. The following relations show that the DNE always decreases the absolute value of eigenvalues,

$$\begin{aligned} 1 &\geq |\lambda|, \\ 1 &\geq \lambda^2, \\ \alpha &\geq \alpha\lambda^2, \\ 1-\alpha &\leq 1-\alpha\lambda^2, \\ |\lambda|(1-\alpha) &\leq |\lambda|(1-\alpha\lambda^2), \\ \frac{|\lambda|(1-\alpha)}{1-\alpha\lambda^2} &\leq |\lambda|, \end{aligned}$$

where the rate of this decrease is higher for eigenvalues with greater absolute values. Figure 12 depicts the behavior of f_α and how this nonlinear function can be regularized with an α parameter. Increasing the eigengaps between large eigenvalues enhances the robustness of the converged network, which in our case means a better formation of classes (for details on the spectral eigengap, see Joseph & Yu, 2016; Wang et al., 2018; Mavroeidis & Bingham, 2010).

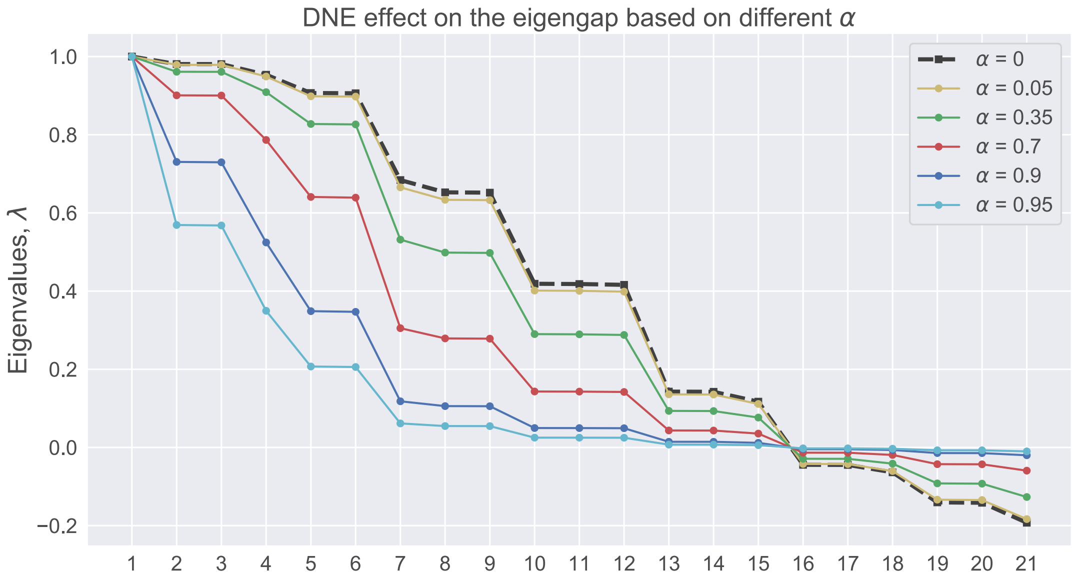


Figure 13: The effect of α on the eigenvalues of the transition matrix of a clip network obtained from experiment 1 in section 4 (see Table 1 for the training structure).

Figure 12 shows that by increasing the regularization parameter, higher eigengaps are achieved. In Figure 13, the associated eigenvalues of a sample network clip[7] and the new eigenvalues of the converged network with different α values are represented.

Appendix B: Training Structure Design Complexity

Here we provide some mathematical calculations to show how complex the design of different training structures could be in real experiments and artificial EPS or E-EPS agents.

Let the set of all classes be $\mathbf{C}$, where each class has m members. Each member of the classes belongs to a separate category, usually labeled by letters A, B, C, and so on. As a result, there are m categories, each with $n = |\mathbf{C}|$ members, so the total number of stimuli equals $m|\mathbf{C}| = mn$. In an arbitrary MTS procedure, the experimenter usually decides how to label categories (among $m!$ possibilities) and which stimuli sets form classes (among $mn!$ possibilities). In real-life experiments, changing the order of two categories (or labels) or how the members of the same class are assembled across different categories might have an impact on the learning and testing outcome.

However, in the computational model, all the categories and stimuli are abstract symbols and are literally the same. We just use the category labels and class indices to differentiate the stimuli. When there is differentiation

[7] The training order is represented in Table 1, and the experiment is clarified in section 4.

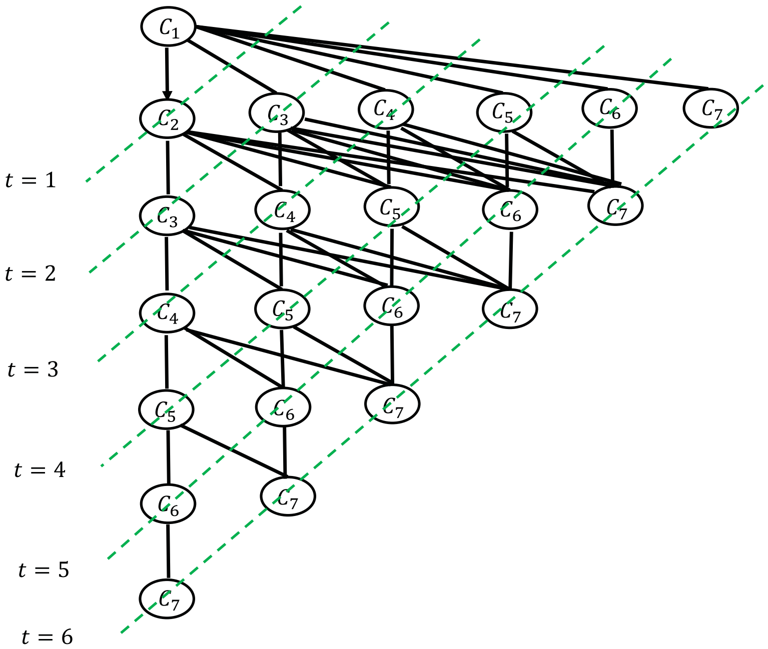


Figure 14: C_1 to C_7 refers to the seven categories and the number of possible maps from categories to C_is, $i = 1, \cdots 7$ is 7!. At each time step, shown by green dashed lines, a category is added to the previously trained relations. At time $t = 1$, the C_1 to C_2 relation, which is shown via a directed connection, is trained as the first relation. This can be any relation. Then at each time step, a new category is connected to the previously trained relations.

between categories in a real-life experiment, the total number of baseline relation configurations, defined as **T**, would be

$$\begin{aligned}\mathbf{T} &= \binom{m}{1}\binom{m-1}{1}\left(2\binom{2}{1}\binom{m-2}{1}\right)\left(2\binom{3}{1}\binom{m-3}{1}\right)\cdots\left(2\binom{m-1}{1}\binom{1}{1}\right)\\ &= 2^{m-2}m!(m-1)! \qquad \text{(B.1)}\end{aligned}$$

In the EPS model, we can remove the repetitions by assuming the category label describes the order of adding a category. For instance, the first relation for training would be AB, the next training could be one of AC, BC, CA or CB, and so on. The number of different training configurations for the agent in this case is

$$\mathbf{T} = 1 \times \left(2\binom{2}{1}\right) \times \left(2\binom{3}{1}\right) \cdots \left(2\binom{m-1}{1}\right) = 2^{m-2}(m-1)! \qquad \text{(B.2)}$$

To make these calculations more intuitive, consider the case with seven categories, that is, $m = 7$, with labels $A, B, C, D, E, F,$ and, G, each with three members $n = 3$. In Figure 14, C_1 to C_7 refers to the seven categories where

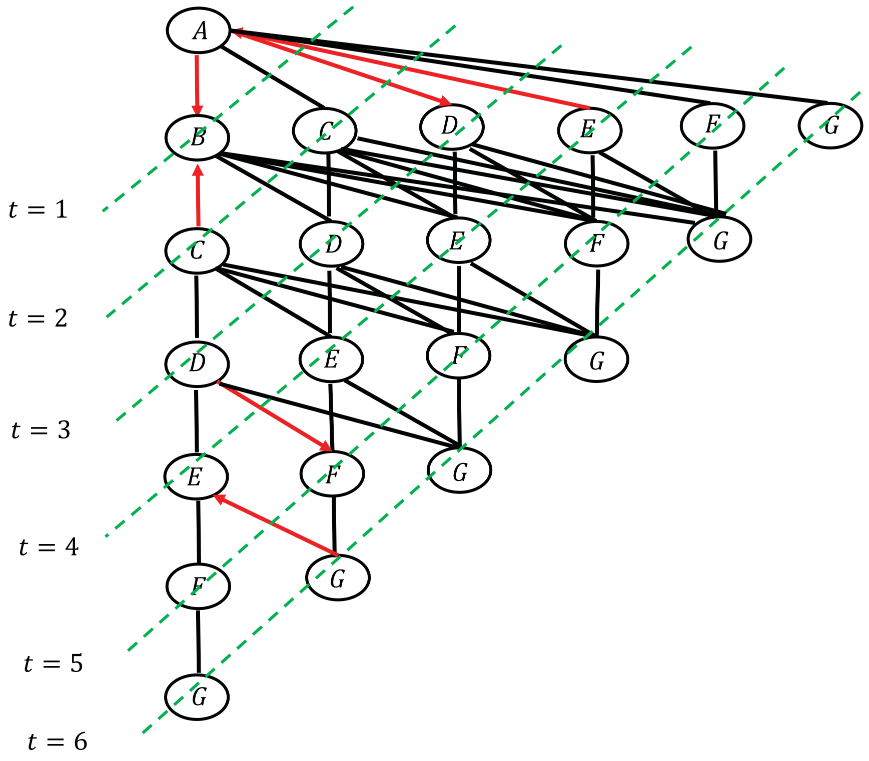
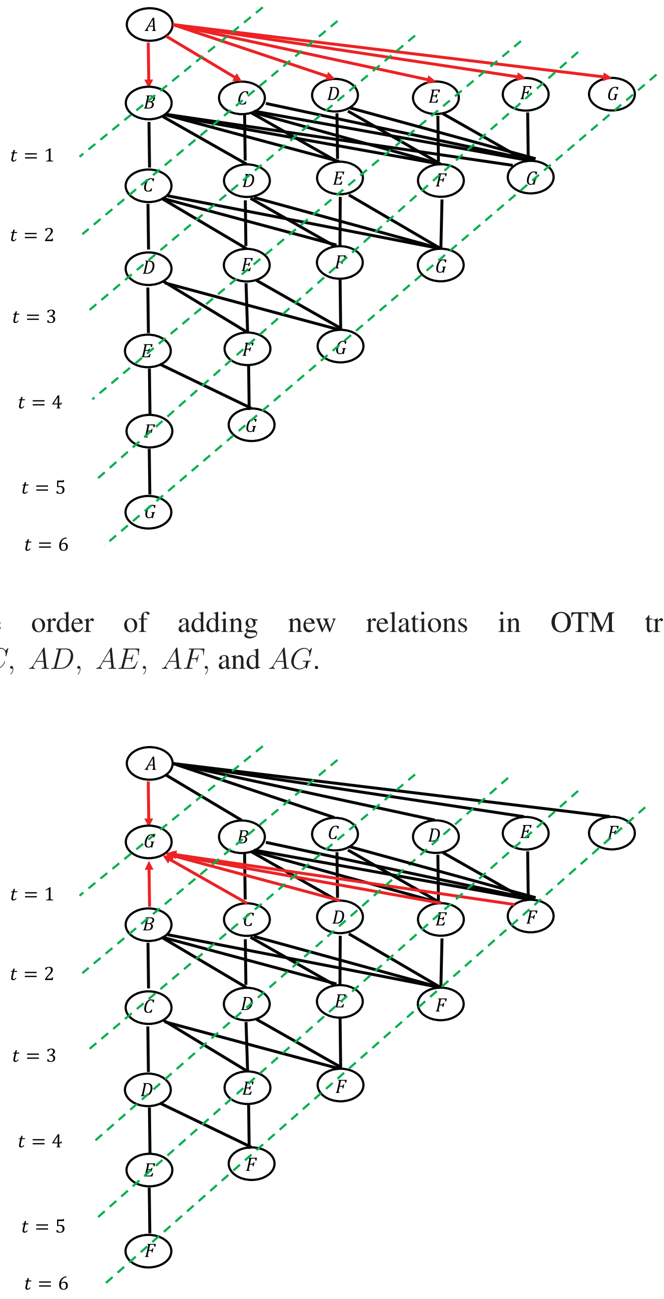


Figure 15: A possible training structure is shown in red—*AB*, *CB*, *AD*, *EA*, *DF*, *GE*—when the order of categories in the training structure is not important.

at each time step, one relation to a new category will be added. The first training stage contains the C_1 to C_2 relation, which is shown via a directed connection. C_1 could be any of seven categories, and C_2 could be one of the remaining six categories. The next stage, represented with $t = 2$ is to add C_3, which is one of the remaining five categories. There are four options to train: C_1C_3, C_3C_1, C_2C_3, and C_3C_2, shown with undirected connections. Similarly, we see that for $t = 3$, there are four choices for categories and 2×3 ways to choose the relation that connects C_4 to previous categories. Therefore, we can easily see that the number of possible maps of categories to C_1 to C_7 is 7! and the possibility them with six relations is $2^5(6!)$. In total, if we distinguish between categories and therefore their order, the number of possible training procedures based on equation B.1 and our explanation equals $2^5(7!)(6!) = 32 \times 5040 \times 720 = 116, 121, 600$.

If we consider the order of categories to be the same and map $C_1 \rightarrow A$, $C_2 \rightarrow B$, $C_3 \rightarrow C$, $C_4 \rightarrow D$, $C_5 \rightarrow E$, $C_6 \rightarrow F$, and $C_7 \rightarrow G$, different configurations will be reduced to $2^5(6!) = 32 \times 720 = 23{,}040$, according to equation B.2. This one-to-one mapping is shown in Figure 15, along with a sample training order in directed red connections that is not LS, OTM, or MTO (see Table 7 for a summary of the training).



(a) The order of adding new relations in OTM training structure: AB, AC, AD, AE, AF, and AG.

(b) The order of adding new relations in MTO training structure: AG, BG, CG, DG, EG, and FG.

Figure 16: Graphical representation of training order for OTM and MTO, shown in red.

In Figures 16a and 16b, respectively, the order of adding new relations to the training blocks for OTM and MTO is depicted. Both training structures are addressed in experiment 1 and reported in Tables 5 and 6.

Although our argument and equations B.1 and B.2 show the complexity of studying the effect of a training structure in an MTS procedure on the

Table 7: Training Order for the Training Structure Depicted in Figure 15.

Time Step	New Relation	Possible Previous Relations				
$t = 1$	AB					
$t = 2$	CB	AB				
$t = 3$	AD	CB	AB			
$t = 4$	EA	AD	CB	AB		
$t = 5$	DF	EA	AD	CB	AB	
$t = 6$	GE	DF	EA	AD	CB	AB

Note: A training block can be formed by only new relation at each stage or a combination of new and previously trained relations.

participant/agent performance, the training structure and training block design are much more complex. We have addressed the order of adding new training relation to the previously trained relations. Many other parameters can be included in the analysis, such as the number of trials in each block, the combination of previously trained relations together with the new relation, testing derived relations during training or not, testing order, and number of classes (members of each category). Moreover, the possibility of training a mixture of relations between two categories, say, A_1B_1, B_2A_1, A_3B_3, will increase this number. An example of such training is simulated in our previous work (Mofrad et al., 2020). Therefore, finding some optimal training structure either theoretically or via simulation with EPS or E-EPS is an interesting problem in its own right, but it is out of the scope of this letter.

Abbreviations

DNE	Directed Network Enhancement.
DSM	doubly stochastic matrix.
E-EPS	Enhanced Equivalence Projective Simulation.
EPS	Equivalence Projective Simulation.
fMRI	Functional Magnetic Resonance Imaging.
LS	linear series.
MTO	many-to-one.
MTS	matching-to-sample.
NE	Network Enhancement.
OTM	one-to-many.
PS	Projective Simulation.
RL	reinforcement learning.
SE	Stimulus Equivalence.
SNE	Symmetric Network Enhancement.

Acknowledgments

We thank the anonymous reviewers for their constructive and thorough feedback that helped us improve the quality of this letter. The source code of the E-EPS model simulator is made available online at https://github.com/Asieh-A-Mofrad/Enhanced-Equivalence-Projective-Simulation.

References

Arntzen, E. (2012). Training and testing parameters in formation of stimulus equivalence: Methodological issues. *European Journal of Behavior Analysis*, *13*(1), 123–135.

Arntzen, E., Grondahl, T., & Eilifsen, C. (2010). The effects of different training structures in the establishment of conditional discriminations and subsequent performance on tests for stimulus equivalence. *Psychological Record*, *60*(3), 437–461.

Arntzen, E., & Hansen, S. (2011). Training structures and the formation of equivalence classes. *European Journal of Behavior Analysis*, *12*(2), 483–503.

Arntzen, E., & Holth, P. (1997). Probability of stimulus equivalence as a function of training design. *Psychological Record*, *47*(2), 309–320.

Arntzen, E., & Mensah, J. (2020). On the effectiveness of including meaningful pictures in the formation of equivalence classes. *Journal of the Experimental Analysis of Behavior*, *113*(2), 305–321.

Barnes, D., & Hampson, P. J. (1993). Stimulus equivalence and connectionism: Implications for behavior analysis and cognitive science. *Psychological Record*, *43*(4), 617–638.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509.

Briegel, H. J., & De las Cuevas, G. (2012). Projective simulation for artificial intelligence. *Scientific Reports*, *2*(1), 1–16.

Cullinan, V. A., Barnes, D., Hampson, P. J., & Lyddy, F. (1994). A transfer of explicitly and nonexplicitly trained sequence responses through equivalence relations: An experimental demonstration and connectionist model. *Psychological Record*, *44*(4), 559–585.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Devany, J. M., Hayes, S. C., & Nelson, R. O. (1986). Equivalence class formation in language-able and language-disabled children. *Journal of the Experimental Analysis of Behavior*, *46*(3), 243–257.

Fields, L., Adams, B. J., Verhave, T., & Newman, S. (1990). The effects of nodality on the formation of equivalence classes. *Journal of the Experimental Analysis of Behavior*, *53*(3), 345–358.

Fienup, D. M., Wright, N. A., & Fields, L. (2015). Optimizing equivalence-based instruction: Effects of training protocols on equivalence class formation. *Journal of Applied Behavior Analysis*, *48*(3), 613–631.

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *Elife*, *6*, e17086.

Groskreutz, N. C., Karsina, A., Miguel, C. F., & Groskreutz, M. P. (2010). Using complex auditory-visual samples to produce emergent relations in children with autism. *Journal of Applied Behavior Analysis*, *43*(1), 131–136.

Hayes, S. C. (1989). Nonhumans have not yet shown stimulus equivalence. *Journal of the Experimental Analysis of Behavior*, *51*(3), 385–392.

Hove, O. (2003). Differential probability of equivalence class formation following a one-to-many versus a many-to-one training structure. *Psychological Record*, *53*(4), 617–634.

Joseph, A., & Yu, B. (2016). Impact of regularization on spectral clustering. *Annals of Statistics*, *44*(4), 1765–1791.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616.

Lew, S. E., & Zanutto, S. B. (2011). A computational theory for the learning of equivalence relations. *Frontiers in Human Neuroscience*, *5*, 113.

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, *8*(3–4), 293–321.

Liu, T.-Y., & Watson, B. O. (2020). Patterned activation of action potential patterns during offline states in the neocortex: Replay and non-replay. *Philosophical Transactions of the Royal Society B*, *375*(1799), 20190233.

Lyddy, F., & Barnes-Holmes, D. (2007). Stimulus equivalence as a function of training protocol in a connectionist network. *Journal of Speech and Language Pathology–Applied Behavior Analysis*, *2*(1), 14.

Lyddy, F., Barnes-Holmes, D., & Hampson, P. J. (2001). A transfer of sequence function via equivalence in a connectionist network. *Psychological Record*, *51*(3), 409–428.

Mautner, J., Makmal, A., Manzano, D., Tiersch, M., & Briegel, H. J. (2015). Projective simulation for classical learning agents: A comprehensive investigation. *New Gener. Comput.*, *33*(1), 69–114.

Mavroeidis, D., & Bingham, E. (2010). Enhancing the stability and efficiency of spectral ordering with partial supervision and feature selection. *Knowledge and Information Systems*, *23*(2), 243–265.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38.

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology*, *4*, 503.

Melnikov, A. A., Makmal, A., Dunjko, V., & Briegel, H. J. (2017). Projective simulation with generalization. *Scientific Reports*, *7*(1), 14430.

Mofrad, A. A., Yazidi, A., Hammer, H. L., & Arntzen, E. (2020). Equivalence projective simulation as a framework for modeling formation of stimulus equivalence classes. *Neural Computation*, *32*(5), 912–968.

Momennejad, I. (2020). Learning structures: Predictive representations, replay, and generalization. *Current Opinion in Behavioral Sciences*, *32*, 155–166.

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2017). *Offline replay supports planning: FMRI evidence from reward revaluation*. bioRxiv:196758.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692.

Ninness, C., Ninness, S. K., Rumph, M., & Lawson, D. (2018). The emergence of stimulus relations: Human and computer learning. *Perspectives on Behavior Science*, *41*(1), 121–154.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

O'Mara, H. (1991). Quantitative and methodological aspects of stimulus equivalence. *Journal of the Experimental Analysis of Behavior*, *55*(1), 125–132.

Parr, T., Markovic, D., Kiebel, S. J., & Friston, K. J. (2019). Neuronal message passing using mean-field, Bethe, and marginal approximations. *Scientific Reports*, *9*(1), 1–18.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Ried, K., Eva, B., Müller, T., & Briegel, H. J. (2019). *How a minimal learning agent can infer the existence of unobserved variables in a complex environment*. arXiv:1910.06985.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLOS Computational Biology*, *13*(9), e1005768.

Schwöbel, S., Kiebel, S., & Marković, D. (2018). Active inference, belief propagation, and the Bethe approximation. *Neural Computation*, *30*(9), 2530–2567.

Shrager, J., Hogg, T., & Huberman, B. A. (1987). Observation of phase transitions in spreading activation networks. *Science*, *236*(4805), 1092–1094.

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech, Language, and Hearing Research*, *14*(1), 5–13.

Sidman, M. (1990). Equivalence relations: Where do they come from? In D. E. Blackman & H. Lejeune (Eds.), *Behaviour analysis in theory and practice: Contributions and controversies* (pp. 93–114). Mahwah, NJ: Erlbaum.

Sidman, M. (1994). *Equivalence relations and behavior: A research story.* Authors Cooperative.

Sidman, M., Cresson Jr., O., & Willson-Morris, M. (1974). Acquisition of matching to sample via mediated transfer 1. *Journal of the Experimental Analysis of Behavior*, *22*(2), 261–273.

Sidman, M., Rauzin, R., Lazar, R., Cunningham, S., Tailby, W., & Carrigan, P. (1982). A search for symmetry in the conditional discriminations of rhesus monkeys, baboons, and children. *Journal of the Experimental Analysis of Behavior*, *37*(1), 23–44.

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, *37*(1), 5–22.

Sidman, M., Willson-Morris, M., & Kirk, B. (1986). Matching-to-sample procedures and the development of equivalence relations: The role of naming. *Analysis and intervention in Developmental Disabilities*, *6*(1–2), 1–19.

Spencer, T. J., & Chase, P. N. (1996). Speed analyses of stimulus equivalence. *Journal of the Experimental Analysis of Behavior*, *65*(3), 643–659.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643.

Steingrimsdottir, H. S., & Arntzen, E. (2011). Using conditional discrimination procedures to study remembering in an Alzheimer′s patient. *Behavioral Interventions*, *26*(3), 179–192.

Stella, F., Baracskay, P., O′Neill, J., & Csicsvari, J. (2019). Hippocampal reactivation of random trajectories resembling Brownian diffusion. *Neuron*, *102*(2), 450–461.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Sutton, R. S., Szepesvári, C., Geramifard, A., & Bowling, M. (2008). Dyna-style planning with linear function approximation and prioritized sweeping. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence* (pp. 528–536). Arlington, VA: AUAI Press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208.

Tovar, Á. E., & Westermann, G. (2017). A neurocomputational approach to trained and transitive relations in equivalence classes. *Frontiers in Psychology*, *8*, 1848.

Wang, B., Pourshafeie, A., Zitnik, M., Zhu, J., Bustamante, C. D., Batzoglou, S., & Leskovec, J. (2018). Network enhancement as a general method to denoise weighted biological networks. *Nature Communications*, *9*(1), 3108.

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science*, *338*(6104), 270–273.

Received March 20, 2020; accepted September 1, 2020.