Explainable Hypergraph Neural Networks for the Prediction of Dementia Progression

John Fletcher

MSc in Computer Science

The University of Bath

2024

Explainable Hypergraph Neural Networks for the Prediction of Dementia Progression

Submitted by: John Fletcher

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of MSc in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Explainable Hypergraph Neural Networks for the Prediction of Dementia Progression

Abstract

Dementia is a progressive illness which affects quality of life for millions of people every year. It has a complex pathology with a variety of associated lifestyle, medical, demographic, genetic and biological factors. Early intervention can delay or mitigate the impact, but early diagnosis is challenging because it depends on observable cognitive decline. To aid clinicians in dementia diagnosis, Machine Learning techniques can be used to predict individual risk from an array of clinical and biomarker features. However, biomarkers are expensive and invasive to collect, posing financial and ethical barriers to their use.

Hypergraphs can model higher-order relationship which other data structures cannot and can use inference from these relationships to classify data. To date, no study has explored the potential of an explainable hypergraph neural network to predict dementia progression using only clinical and genetic features without other biomarkers. In this work, we fulfil this objective with a cutting-edge Equivariant Hypergraph Neural Network (EHNN) augmented with a self-explainable module to observe the model's reasoning.

Using EHNN with cost-sensitive learning to predict dementia progression over 3 years, we achieve an F_1 score of 0.73, 79% accuracy, an AUC_{ROC} of 0.86 and an AUC_{ROC} of 0.75 (against baseline 0.34), showing good skill distinguishing positive and negative classes while minimizing false negatives, both clinically important capabilities. We incorporate a factual and counter-factual self-explainability module and demonstrate that explainability does not have to come at the cost of performance. Indeed, not only does this maintain the F_1 score of 0.73, but also improves AUC_{ROC} to 0.87 and AUC_{ROC} to 0.76.

Finally, we perform high level analysis on the explainable outputs and find that current cognitive test scores are a good indicator of future dementia progression, agreeing with existing literature. We find some surprising results, that APOE4 and family history of dementia do not correlate with dementia prediction in the model's reasoning; however, further evidence shows that this is likely a misinterpretation of the outputs and that instead, the model uses this information as a signal to concentrate on other risk factors.

This study finds that Explainable Hypergraph Neural Networks can provide clinically meaningful results both in predicting dementia progression with low-cost, non-invasive features and in explaining dementia risk factors in large groups. However, the underlying data contains many biases and additional work is needed in explaining the model's outputs before such a model can be recommended for clinical deployment.

Contents

List of Figures

List of Tables

List of Acronyms

Aβ	Amyloid-Beta
ACS	Alzheimer's Clinical Syndrome
AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
APOE4	Gene encoding for Apolipoprotein E4
AUC_{PR}	Area under the Precision-Recall Curve
AUC_{ROC}	Area under the Receiver Operating Curve
BMI	Body Mass Index
BCE	Binary Cross Entropy
CDR	Clinical Dementia Rating
CDR-GS	Clinical Dementia Rating – Global Score
CDR-SB	Clinical Dementia Rating – Sum of Boxes
CSF	Cerebrospinal Fluid
DSM	Diagnostic and Statistical Manual of Mental Disorders
EDA	Exploratory Data Analysis
EHNN	Equivariant Hypergraph Neural Network
GPU	Graphics Processing Unit
HIV	Human Immunodeficiency Virus
ML	Machine Learning
MLP	Multi-Layered Perceptron
MMSE	Mini Mental State Examination
NACC	National Alzheimer's Coordinating Center
NIA-AA	National Institute on Aging – Alzheimer's Association
NHS	National Health Service (United Kingdom)
OASIS	Open Access Series of Imaging Studies
PET	Positron Emission Tomography
RQ	Research Question
SVM	Support Vector Machine
TF_1	Test set F_1 score
VF_1	Validation set F_1 score
WHO	World Health Organisation

Acknowledgements

I would like to thank my dissertation advisor Dr. Alok Joshi for his guidance towards this topic and continuous support in improving and refining the work. I thank my wife Tassanee and my family for their support which was invaluable to completing this project while also working in a full-time job.

Chapter 1

Introduction

This chapter introduces the problem under investigation and existing challenges faced in addressing it. We then discuss specific objectives in relation to the problem, contributions to the field and the overall structure of the dissertation.

1.1 Problem Space: Predicting Dementia Progression

According to the World Health Organisation (WHO, 2023), more than fifty-five million people suffer from dementia with sixty percent living in low- and middle-income countries. The aetiology of the disease is complex, and there are many sub-types. There is no cure, but early interventions can alleviate symptoms (Livingston et al., 2017). Therefore, predicting dementia progression can lead to improved outcomes.

Dementia progression prediction is a growing research field; however, studies have mainly focused on expensive and invasive biomarkers; the majority of publications in the last two decades focus on neuroimaging as the key modality, particularly Positron Emission Topology (PET) (Battineni et al., 2022). Wittenberg et al, (2019) estimate it would cost £113 million to introduce 100,000 such scans in the UK National Health Service (NHS). A tracer fluid consumed for each scan accounts for eighty percent of the cost, so there is little economy of scale. Around twenty percent of the UK population, 13.5 million people, are over the age of sixty-five with seven percent affected by the most common dementia sub-type Alzheimer's Disease (AD) (NHS, 2018). It is, therefore, not financially feasible to scan every person over the age 65.

Clinical and non-invasive markers are cheaper to collect: Genetic testing kits for AD can be privately purchased in the UK for £149 (Alzheimer's Society, n.d.) and many clinical markers are already present in medical records or can be recorded though patient assessments. They can tell us a lot about an individual's risk: genetic profile, education level, hearing loss, hypertension and obesity together contribute twenty-seven percent of known early- and mid-life dementia risk while smoking, depression, physical inactivity, social isolation and diabetes contribute fifteen percent of late-life risk (Livingston et al., 2017). Further, there are complex interrelations between risks, demographics, and habits which may provide additional insight. For example, Yuan et al (2022) found a negative correlation between high BMI (Body Mass Index) and AD

risk as individuals age and Fitzhugh and Pa (2022) found that women with hearing loss have a higher risk than men with hearing loss. These clinical characteristics can provide significant insight into dementia risk without the need for more expensive and invasive investigations.

We hypothesize that dementia risk prediction with only clinical markers can be improved by learning from the complex interrelationships between them. Existing studies using only clinical factors have been limited to modeling data in pairwise relationships between risk factor and disease probability (Rowe et al., 2021; Battineni et al., 2022). The key disadvantage of these models is an inability to account for polyadic relationships between risk factors in a population.

Hypergraph neural networks are new Machine Learning (ML) techniques which take advantage of a hypergraph's structure to model higher order relationships in data. A hypergraph allows multiple nodes to be connected with a single hyperedge whereas a graph edge can only join two nodes (Torres et al., 2020). A hyperedge can thus model complex group relationships not possible with other data structures. **Figure 1** illustrates a simplified example for modeling dementia risk factors with a graph compared to a hypergraph. In this example, unlike the hypergraph, the graph is unable to capture the incidence of APOE4 alleles as a three-way group relationship. This study aims to leverage a hypergraph neural network for dementia progression prediction where individuals are represented as nodes and risk factors as hyperedges.

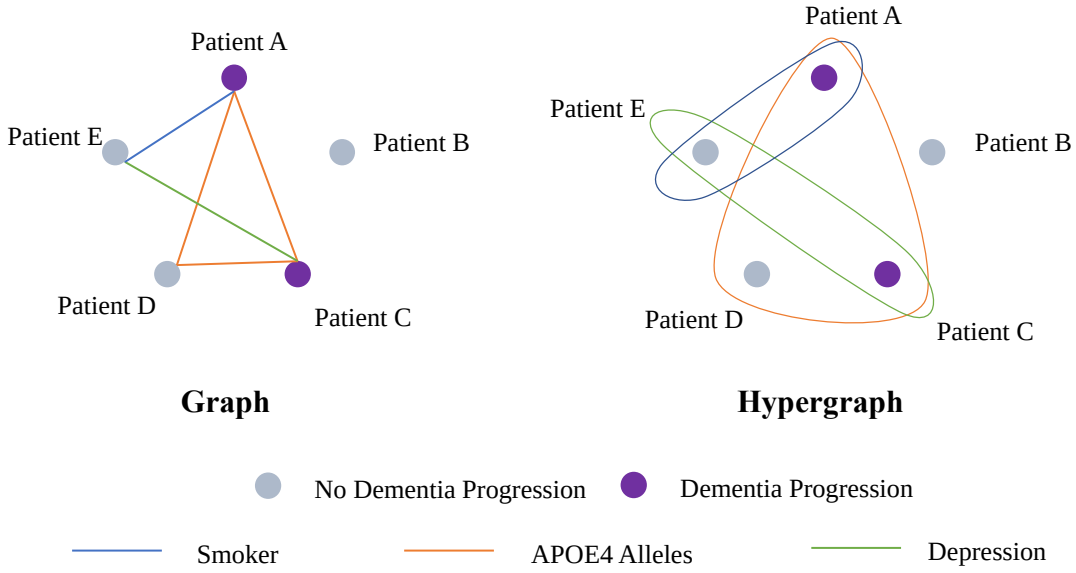

Figure 1: *Graph vs Hypergraph.* Both the graph and hypergraph model the same data with patients as nodes, which can be classified as having progressed to dementia or not, and risk factors as (hyper)edges. Unlike the hypergraph, the graph cannot capture the higher order relationship that Patient A, C and D share the APOE4 alleles feature in a three-way relationship.

In addition to implementing a hypergraph neural network, this project aims to address another gap in dementia progression prediction: explainability. Many such studies are designed without consideration to clinical relevance or introspection (Ansart et al., 2021). We tackle this issue by carefully defining our model's goal in clinical terms and implementing an explainability module. Further, we aim to show that this is possible without harming performance, a common argument against using explainability techniques (Hatherley, Sparrow and Howard, 2022).

1.2 Project Aims

This exploratory project aims to apply a state-of-the-art hypergraph neural network model to predict progression of dementia using non-invasive, clinical and genetic features and to apply a self-explainability module. We will explore the use of the explainability module's output as a means to understanding the model's reasoning.

1.3 Objectives

- Create a hypergraph structure with clinical and genetic data for the prediction of dementia progression.
- Train a hypergraph neural network to classify at-risk individuals for dementia progression with this data.
- Modify the hypergraph neural network with a factual and counter-factual reasoning module.
- Analyse the causal features identified by the hypergraph neural network.
- Identify limitations of the data and model in order to make recommendations for future studies.

1.4 Novel Contributions

- Show hypergraph neural networks can outperform existing methods for dementia progression prediction with clinical and genetic features.
- Show we can maintain dementia progression risk prediction performance with the hypergraph neural network when including a self-explainability module.
- Assess the explainability-augmented hypergraph neural network's limitations by examining the self-explainability module's explanations for its dementia risk classifications.

1.5 Structure

This dissertation will aim to fulfill these goals by following this structure:

Chapter 1 – Introduction

This chapter outlines, at a high level, the problem of dementia progression prediction, challenges in tackling it and the gap which this research fills. It describes the project's objectives and novel contributions and the general structure of the dissertation.

Chapter 2 – Literature and Technology Review

In this chapter we provide background into the main topics which need to be understood to address the objectives: the clinical and research contexts for dementia and associated risk factors, the state of current literature on dementia progression prediction and a brief overview of hypergraph neural networks.

Chapter 3 – Methodology, Data and Model

This chapter sets out our specific research questions and how we will answer them. We first explain the overall process, then explain the choice of dataset, analyse it to understand its high-level shape and features before transforming it appropriately for a hypergraph neural network. We then discuss the specifics of the hypergraph neural network and explainability techniques we employ in our experiments, ending with a discussion on measuring model performance.

Chapter 4 – Experiments, Results, and Analysis

In this chapter we conduct a series of experiments with our hypergraph neural network model and explainability module demonstrating each contribution in turn. We first assess model performance with different sub-datasets and model hyper-parameters without the explainability module. We then implement the self-explainability module and observe its impact on model performance across a range of hyper-parameters, comparing the best result against those achieved without the module. Finally, we perform high level analysis on the module's explanatory outputs.

Chapter 5 – Conclusions

This chapter provides a summary of our findings and contributions alongside limitations and suggestions for future work to address them.

Chapter 2

Literature and Technology Review

This chapter provides background into this project's main topics. We first contrast how dementia is classified in clinical and research contexts, explaining why the former is more relevant for this project. Next, we explore some known risk factors and their role in clinical assessment. We then discuss the existing state of the literature for dementia progression prediction and introduce hypergraph neural networks as a classification tool. We end with a discussion on explainability and its clinical relevance.

2.1 Defining Dementia for Classification

To classify dementia, we must be able to first define it, and to predict progression of dementia, we need some understanding of the associated causes and risk factors. In this section, we explore dementia types and compare two contexts in which dementia is commonly defined: clinical and research, illustrated in **Figure 2**. We explain why we focus on a clinical definition and the existing clinical challenges with assessing dementia risk.

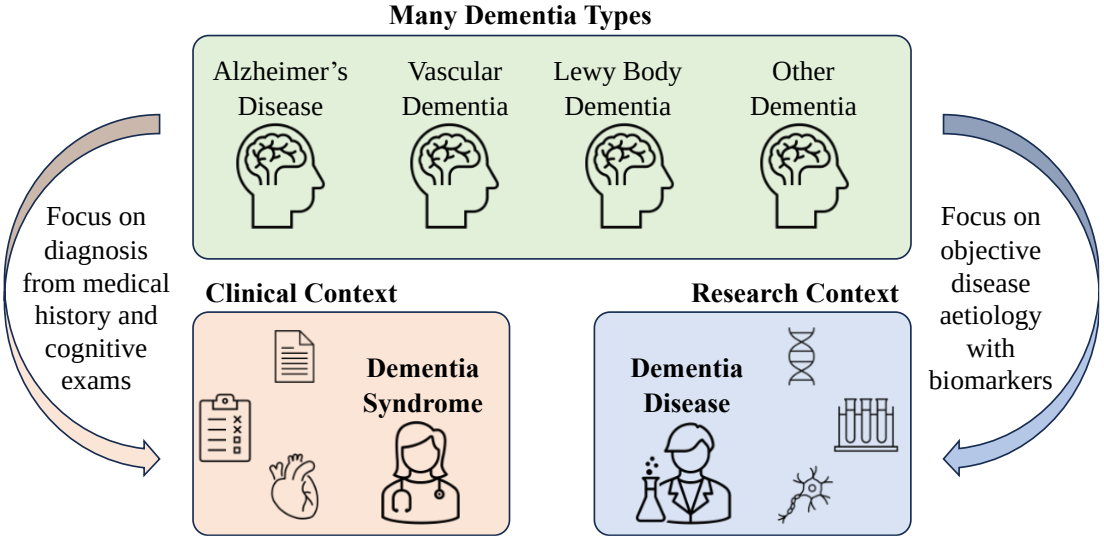

Figure 2: *Dementia Types and Contexts.* Dementia is an umbrella term for several illnesses which manifest in cognitive decline. Clinical focus is on diagnosis and treatment of dementia syndrome. Research focuses on the precise physiological aetiology of dementia disease.

2.1.1 Types of dementia

Dementia first appeared as a clinical diagnosis in the 18th century characterized only by observable behavioral changes, but by the end of the 19th century physiological links, such as temporal lobe atrophy, were established (Burns and Levy, 1994, pp.5–15). Today, dementia is an umbrella term for a variety of specific illnesses with differing causes but manifesting in similar symptoms. Each type has a complicated, long-term and cascading pathology which varies greatly between individuals (Barnes and Lee, 2011; Tahami Monfared et al., 2022). The three most common types of dementia, in order, are: Alzheimer's Disease, affecting fifty to eighty percent of cases and caused primarily by changes in amyloid-beta plaque buildup in the brain; Vascular Dementia, affecting up to thirty percent of cases and caused by vascular problems such as heart disease and strokes; and Lewy Body Dementia, caused by buildup of alpha-synuclein in the brain and closely associated with Parkinson's disease (Hobson, 2019, pp.30–35; Livingston et al., 2017). There are also forms of dementia associated with co-morbidities such as HIV (human immunodeficiency virus), alcohol abuse and traumatic brain injury but they are less common (Livingston et al., 2017).

2.1.2 Diagnostic contexts for dementia

There are two contexts in which dementia may have different diagnostic definitions: the clinical context, and the research context. The clinical context focuses on diagnosis and treatment within a subjective clinical setting whereas the research context seeks objective aetiological explanations which may be used for practical applications. In this section, we explore both contexts in relation to our research question.

2.1.2.1 Dementia diagnosis in a clinical context

In a clinical setting, dementia is primarily diagnosed through the use of cognition tests which are defined in frameworks such as the DSM (Diagnostic and Statistical Manual of Mental Disorders) and MMSE (Mini Mental State Examination) (Pelegrini et al., 2019; Pais et al., 2020). Since pathology can manifest long before cognitive and behavioral changes (Van Der Schaar et al., 2022), a clinical diagnosis is likely to occur after the optimal time for interventions. While there are biomarkers, such as amyloid-beta and tau protein levels, which can indicate potential dementia onset before external symptoms manifest, their preemptive use raises ethical and financial questions since they are invasive and expensive to collect (Ford, Milne and Curlewis, 2023). Further, Van Der Schaar et al (2022, p.8) highlight that the majority of cognitively healthy individuals who have abnormal biomarkers do not develop dementia. Therefore, the clinical approach to assessing dementia focuses on the cheaper, externally measurable, and less ethically ambiguous signs of dementia.

2.1.2.2 Dementia diagnosis in a research context

When conducting aetiological research, the National Institute on Aging – Alzheimer's Association (NIA-AA) strongly encourages researchers to focus on a biological classification of AD with biomarkers and when deciding to use clinical features, to clarify research as considering "Alzheimer's clinical syndrome" (ACS) rather than AD (Jack et al., 2018). To demonstrate the need for this distinction, Jack et al (2018, p.21) cite a specific example where an individual was incorrectly diagnosed with AD by several physicians but biomarkers later revealed a diagnosis of non-AD pathological changes. Further, Jack et al (2018) highlight the risk of conflating the pathology of general dementia co-morbidities with AD, citing a non-biomarker study which identified diabetes as a risk factor for AD which was later discovered in an autopsy to be pathologically associated to vascular brain injury, a different dementia type. Therefore, when designing our study, we must make a measured decision to use either a symptom-based or biomarker-based definition of dementia.

2.1.2.3 Defining dementia in our study

Although our study is experimental, our results will contribute towards the field of clinical decision support for dementia progression prediction. This is fundamentally a clinical context where a physician faces practical and ethical constraints in using biomarkers. To be relevant in this context, our model should, therefore, be defined with the same limits. We must also avoid attempting to make precise sub-type diagnoses which, as mentioned in the previous section, can be pathologically conflated. Thus, in our study, we choose a clinical definition of "Dementia clinical syndrome".

2.1.3 Dementia risk factors

In early stages of impairment, or even before impairment manifests, an understanding of risk factors allows clinicians to advise patients on potential lifestyle interventions to reduce risk or delay decline and may allow for early pharmaceutical intervention. Indeed, drugs such as Donanemab have been found to delay early stage cognitive decline for AD suffers by reducing amyloid-beta buildup (Mintun et al., 2021).

Assessing risk, however, is not straight-forward. There are many factors and every individual's profile is complex and unique; Livingston et al, (2017) highlight a wide array of social, medical and lifestyle factors which contribute. Further, co-morbidities do not share the same relationships with all types of dementia at all stages. For example, Gerritsen et al (2016) found those suffering from young-onset AD were less likely than late-onset sufferers to have co-morbidities such as diabetes but were more likely to have diseases of the nervous system. Some have a more pronounced effect for specific genders, such as hearing loss for women (Fitzhugh and Pa, 2022), or have

complex interrelationships between genes and other risk factors (Yuan et al., 2022). Predicting overall risk, therefore, requires an ability to understand and account for the specific characteristics and relationships for an individual. Therefore, it is highly desirable to improve the clinical decision-making process for dementia risk and diagnosis.

2.2 Predicting Dementia Progression with ML

Given the challenging but valuable nature of forecasting risk, there has been an increasing effort to develop ML solutions over the last twenty years (Ansart et al., 2021). Experimental systems have been developed using a range of statistical ML models and neural networks. In a survey of articles aiming to predict dementia or AD specifically, Kumar et al (2021) found Support Vector Machines SVMs to be the most popular technique, followed by neural networks of varying types, and about half considered only clinical data rather than clinical and imaging data. Performance of models varies, particularly depending on the input feature sets and the specific definition of the classification task.

Hypergraphs have been used to study AD risk, but not for clinical risk factors. Wang et al (2022), Zuo et al (2021), and Shao et al (2020) have used hypergraphs to model brain structure relationships in neuroimages, Shao et al (2021) examined the relationships between different genetic markers by comparing hypergraph constructions, Aviles-Rivero et al (2022) combined neuroimaging, age and genetic data into a multi-modal hypergraph and Zuo et al (2021) constructed multi-view hypergraphs from PET, MRI (Magnetic Resonance Imaging) and CSF (Cerebrospinal Fluid) data. These studies have shown promising improvements over existing techniques, but none consider the wider range of clinical risk factors and most depend on imaging as the only modality. Hypergraphs have been used in other domains for prediction tasks with large heterogenous datasets. For example, Li et al (2022) successfully created a hypergraph for node classification of student performance using behaviours – analogous to risk factors – as hyperedges. Furthermore, the model also provided useful information on which behaviours were most likely to increase student performance. Therefore, the use of hypergraphs for epidemiological analysis on clinical data is an exciting and potentially powerful tool meriting study.

2.3 Hypergraph Neural Networks for Classification

To overcome data correlation limitations of graph neural networks, Feng et al (2019) introduced the first hypergraph neural network framework using spectral convolution over a hypergraph on a Fourier basis. This model has since been built upon with notable advancements from Bai et al (2020) who developed an attention mechanism to learn a dynamic incidence matrix which allows for richer convolution by modelling

transition probability in parallel to convolution and Chien et al (2021) who generalised the propagation of vertex to edge and edge to vertex learning as special cases of two multi-set functions learned through a set transformer.

In this study, we use a model called Equivariant Hypergraph Neural Network (EHNN) (Kim et al., 2022) which advances the set transformer approach by modelling hypergraphs as a sequence of higher-order tensors representing fully expressive linear layers which share trainable parameters through hypernetworks. The adaptation of this model is discussed further in **section 3.8**.

2.4 Explaining ML Model Predictions

While the overall performance of a model is important, there are also significant socio-technical challenges. Such systems must be both reliable and transparent for clinicians, giving them the most possible data to inform their decisions since a diagnosis of early onset dementia is not always clearly beneficial to patients (Dubois et al., 2015). The diagnosis itself can lead to anxiety, discrimination or stigmatization and incorrect diagnoses may lead to the prescription of unnecessary medication and treatments.

However, deep learning models are often inherently black box; the complexity which allows them to solve difficult problems also makes their reasoning uninterpretable. Further, many AI systems are published without details of how they work or what data they were trained on, which prohibits introspection (Rajpurkar et al., 2022). For low-stakes decisions this may be acceptable, but in a clinical setting this could lead to mistrust of model outputs, especially if they are unexpected (European Commission, 2021). Moreover, this lack of transparency could hide an innovation discovered by the model which would ideally be uncovered and subject to further study (Rajpurkar et al., 2022). The potential for systematic bias is also exacerbated by a lack of transparency. Simply training a model on a large dataset does not naturally result in fair outcomes (European Commission, 2021). To address these issues, we explore the use of an explainability module in addition to the base EHNN model.

In graph learning, explainability methods can be grouped as factual or counterfactual, and post-hoc or self-interpretable (Kakkad et al., 2023). Factual methods identify nodes with the most influence over prediction while counter-factual methods learn by finding variations in the input graph which change the prediction. Post-hoc methods, as shown in **Figure 3**, attempt to explain the reasoning of a model after it has been trained. This has significant disadvantages: when developed separate to the model in question, the companion model can only provide an approximate explanation and may be inaccurate, provide explanations that are not detailed enough or are themselves overly complex, and may show explanations which seem convincing for one specific output but which are in reality very similar for all other outputs (Hatherley, Sparrow

and Howard, 2022). Self-interpretable techniques, on the other hand, include a module which is trained alongside the underlying model to minimise a joint loss function. Explainability is embedded directly in the model's architecture and so provides an authentic view of the model's reasoning processes. Therefore, we select a self-interpretable technique which employs both factual and counter-factual reasoning.

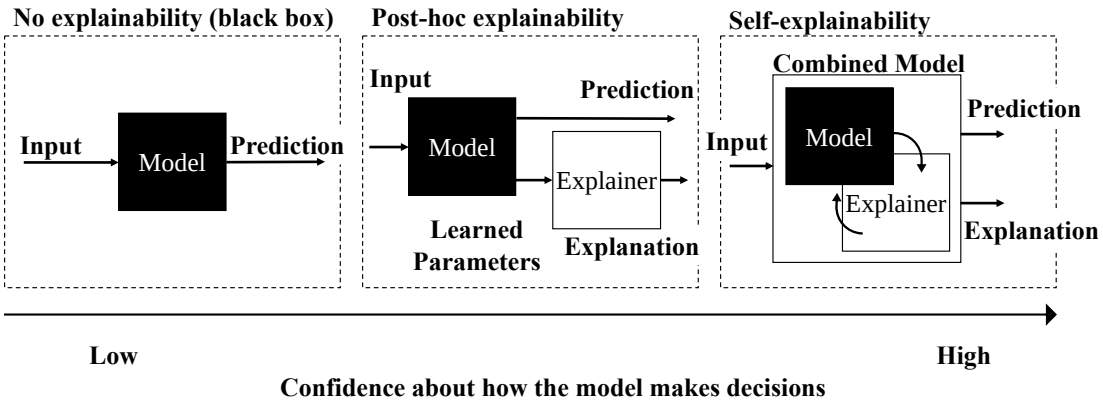

Figure 3: *Confidence and Explainability in ML Models.* Self-explainability gives the highest confidence in decisions compared to post-hoc methods and complete black box methods.

There is an argument against explainability if it comes at the cost of performance, particularly in the medical domain where it poses an ethical dilemma of providing worse but more explainable results (Hatherley, Sparrow and Howard, 2022). We will evaluate this in our study by comparing results with and without explainability.

Xu et al (2022) developed a self-explainability technique for hypergraph learning. They use a subset learner to dynamically select the most important nodes for edge classification by finding factual subsets which generate the same prediction as using the whole graph and counter-factual subsets which produce different predictions with the inverse set of nodes. This technique can be modified for use with any hypergraph neural network and will be employed in this study to achieve explainability.

Dementia is a complex illness and pre-emptive diagnosis is a challenging task. As a result, prediction with ML is a growing field; however, there is a significant gap in the literature where an explainable hypergraph neural network's capability to learn from higher order relationships may be used to improve prediction. We will address this gap by modifying the EHNN model with an explainability module and training it on a large set of clinical and genetic data. The details of this method will be set out in the following chapter.

Chapter 3

Methodology

This chapter explains how we will achieve our goals. We first phrase our objectives as research questions, then describe how a typical ML project pipeline addresses such questions. We use this context to frame them in a clinically relevant manner. Next, we select and explore the data in depth and end the chapter by explaining the technical details of the model and evaluation methods which we will employ in our experiments.

3.1 Research Questions

1. *RQ1:* Can a hypergraph neural network outperform existing methods for dementia progression prediction with clinical and genetic features?
2. *RQ2:* Can we maintain dementia progression risk prediction performance with the hypergraph neural network when including a self-explainability module?
3. *RQ3:* Can we assess the explainability-augmented hypergraph neural network's limitations by examining the self-explainability module's explanations for its dementia risk classification?

3.2 Project Pipeline Overview

A typical ML project framework includes: Framing the problem, preparing the data, choosing and training a model, evaluation, disseminating results, and deploying the model (Panesar, 2020). **Figure 4** details the high-level steps we follow. First, we frame the classification goal based on a literature review of the problem space enabling us to set parameters within which we can select and process data, define our model and assess performance. Then we select, explore, and process a suitable data set. Finally, we choose a state-of-the-art hypergraph model, adapt it for experimentation, and analyse the results. As this is an experimental project, the model will not be deployed.

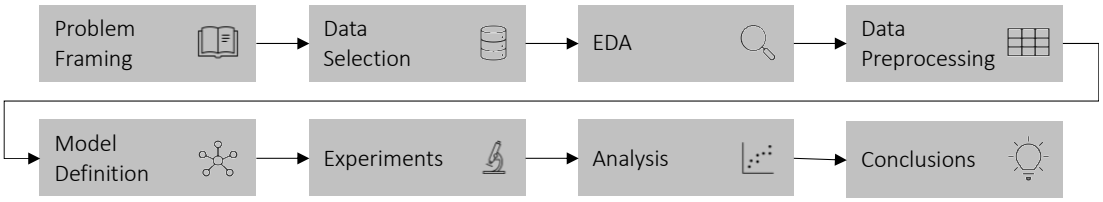

Figure 4: *ML Project Pipeline Overview.* Our project process, a typical ML project.

3.3 Problem Framing

To train a clinically relevant ML model, we need a precisely defined problem statement with two key components: classifying dementia status and measuring progression over a period of time. Unfortunately, these measures are often under considered in research, leading to discrepancies in findings and undermining reproducibility and validity (Marcos et al., 2006; Jack et al., 2018; Ansart et al., 2021). To avoid these pitfalls, we examine the clinical and research background to these definitions before selecting measure for this study.

3.3.1 Classifying dementia conditions

As discussed in section **2.1.2.3.** we will use the "Clinical Dementia Syndrome" definition and so consider only clinical measures to select and classify individuals. Further, physicians generally need to determine progression risk for patients who have some early indication of dementia rather than for those who are cognitively normal (Ansart et al., 2021). Applying these constraints contextualizes our model for clinical applications. Therefore, we look to existing clinical measures.

Most datasets include measures for two standardized clinical cognition tests: CDR (Clinical Dementia Rating) and MMSE. MMSE has been shown to discriminate well between mild cognitive impairment and later stages of dementia while CDR can be used to more accurately differentiate normal cognition and early cognitive impairment (Perneczky et al., 2006). Therefore, we select CDR in this study as a more reliable early screening measure.

CDR scores impairment on a five point scale across multiple categories[1] from "none (CDR = 0)" to "severe (CDR = 3)" (Burns and Levy, 1994, p.358). Scores are then weighted and averaged with the heaviest weight on the memory section. Each category is completed by a clinician through semi-structured interviews with the subject and a co-participant (Burns and Levy, 1994, p.358; NACC, 2015). The NACC dataset includes two summary scores: CDR-GS (global score) and CDR-SB (Sum of boxes). Perneczky et al (2006) found that CDR-GS has a higher pooled sensitivity and specificity (0.99 and 1) than CDR-SB (0.87 and 0.94) and caution against the use of CDR-SB as a screening tool. Further, post-mortem studies (Morris et al, 1991 in Burns and Levy, 1994, p.358; Saito and Murayama, 2007) found that a CDR score of 0.5 ("questionable") is a reliable indicator of neurophysiological deterioration consistent with, but not conclusive of, dementia which is not present in subjects with a CDR of

[1] The full table can be found in **Appendix A**

zero. A CDR-GS score of one demonstrates high probability of mild dementia and so can be considered as dementia progression.

Therefore, this study considers a CDR-GS score of 0.5 as the most relevant early indicator of dementia risk available in datasets. We will use this to select participants for our study as a proxy for a clinical decision to apply early screening and consider a score of at least one to be equivalent to a clinical diagnosis of dementia progression.

3.3.2 Measuring dementia progression over time

We can consider progression as having three components: a starting condition, a progression condition, and a timeframe within which progression occurs. Having already defined the first two, we now consider the progression period.

We first rule out attempting to predict time to progression as this approach risks bias towards those with a longer time under study and is not clinically relevant since it requires foreknowledge of dementia progression status (Ansart et al., 2021). It is, therefore, more clinically relevant, to consider only likelihood of progression over a set period. Ansart et al (2021) found through a meta-study that a minimum of three years is required between initial visit and prediction timeframe to achieve reliable results. In this study, we will experiment with prediction over periods from one to ten years and assess the impact of prediction period on performance.

We can now define a specific, clinically relevant problem statement for our model:

"Will an individual progress from a CDR-GS score of 0.5 to a score equal to or greater than 1 within a fixed time period?"

Equipped with this precise question, we now select a dataset which gives us the greatest breadth of clinical information to train a model to answer it.

3.4 Public Dataset Selection

There are multiple available datasets with longitudinal data on individuals suffering or at risk of dementia. To maximize model performance, we consider the advantages and disadvantages of these datasets before selecting one.

The ADNI (Alzheimer's Disease Neuroimaging Initiative), OASIS (Open Access Series of Imaging Studies), UK Biobank and NACC (National Alzheimer's Coordinating Center) datasets were considered for this research. Although the UK biobank has the largest number of participants (around half a million), it does not focus on dementia nor does it utilize a standardized cognition test (Fawns-Ritchie and Deary, 2020) and so was ruled out as it cannot be used to reliably and consistently assess

cognitive function. The ADNI and OASIS datasets focus primarily on imaging and have smaller cohort sizes of around one to two thousand participants total. The NACC dataset has a participant set of over 33,000 with at least a one-year period recorded. It also includes multiple, low-cost, and non-invasive features for most participants.

We therefore select the NACC dataset and consider NACC participants as eligible for each prediction period if they have the requisite number of visits after being assessed as having a CDR-GS of 0.5. With a chosen dataset, we now examine this data for biases, imbalance and anomalies in an EDA (Exploratory Data Analysis) task.

3.5 Exploratory Data Analysis

Exploratory Data Analysis is defined by (MacInnes, 2020, p.2) as:

> *"an approach to statistics that stresses the importance of the researcher having a good knowledge of how their data were produced, of carefully studying and visualizing the data in order to understand its structure"*

It allows us to gain an initial understanding of the NACC data, surface relationships, errors or biases and informs data remediation (Chandramouli, Dutt and Das, 2018). We will perform EDA in a broad context by studying patterns across different cohorts distinguished by length of participation before diving deeper into specific features of a single period cohort. Data has been visualized with the matlibplot Python package.

3.5.1 Trends in participation-length cohorts

NACC subjects undertake examinations on a yearly basis. These time periods can be used to group participants into participation-length cohorts and observe dementia progression within those groups.

New participants regularly enter the NACC program, so there is a wide range of total visit numbers. To understand trade-offs between total sample size and prediction period, we plot the number of participants by the cumulative number of visits in **Figure 5**. We observe a steady decrease in size as years of participation increase, which is unsurprising as new participants enter and existing participants are lost to attrition or death. This shows us that as we increase the prediction period, we must learn from a smaller input set.

We next examine progression rates by cohort in **Figure 6**. Progression rate increases to a peak at 6 years and then gradually declines, possibly due to a survivor bias over longer periods. Since we are classifying progression and it is not a fifty-fifty split in any cohort, this shows that the data is imbalanced. This is a typical problem in medical

ML due to the nature of diagnosing conditions within an otherwise healthy population and can bias the model if it learns to minimise loss by focusing on the majority class (Fernández et al., 2018; Basit et al., 2022). Therefore, we can infer that training a model on the shorter and longer participation cohorts may result in poor performance.

Since dementia is primarily a disease affecting later life, age differences between cohorts may introduce biases. To anticipate this, we examine age distribution at the initial visit for each cohort in **Figure 7**. All cohorts have a similar distribution with an interquartile age range between 65 and 80; most subjects have an advanced age on their first visit. The slight decrease in age over participation length is potentially a sign of attrition or survivor bias. We also note that all cohorts have age outliers, particularly in shorter participation length cohorts which may adversely affect model performance by obfuscating more general trends (Joshi, 2023, p.84).

By examining macro trends across different length-of-participation cohorts, we have identified three important considerations for data pre-processing and model assessment. First, there is a trade-off between length of prediction period and the total available data to train a model. Second, shorter and longer participation-length cohorts have the greatest class imbalances, likely worsening performance. Third, the data has age outliers which should be removed.

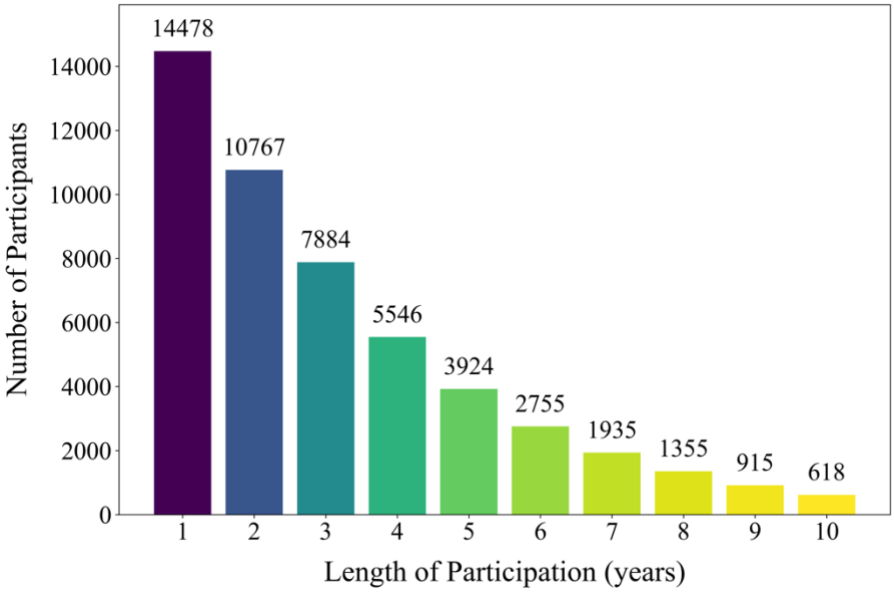

Figure 5: *NACC Cohort Size by Length of Participation.* Shows cumulative number of participants with at least *x* years of NACC visits without any filtering for missing data. As length of participation increases, cohort size decreases.

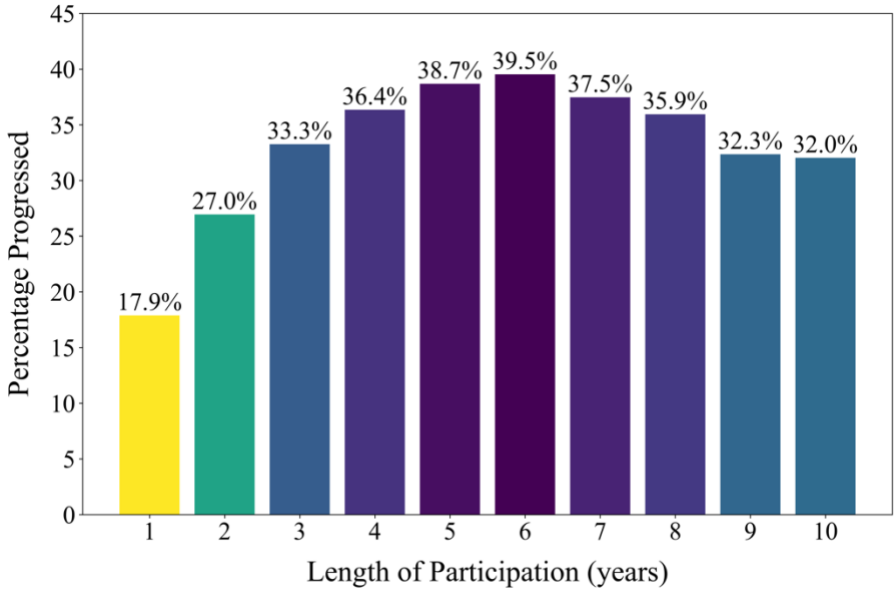

Figure 6: *CDR Score Progression by Length of Participation.* Shows the percentage of participants in each cohort who progressed to a dementia classification. There is a steady increase up to 6 years followed by a decrease. The increase is expected as participant age increases, the decrease may be explained by a survivor bias.

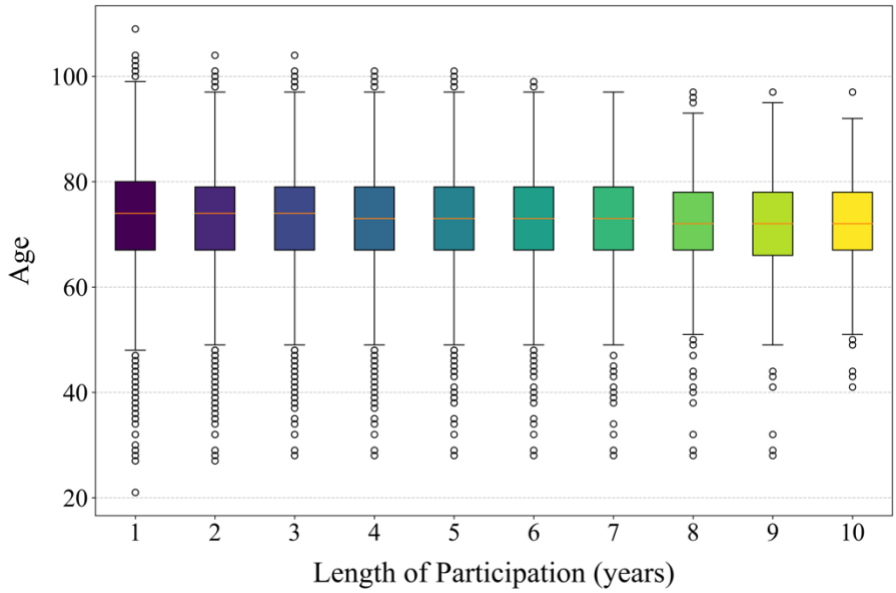

Figure 7: *Initial Age Distribution by Length of Participation.* A box and whisker plot of age on first visit for each cohort. There is a slight gradual decrease in median initial age as length of participation increases which is an expected consequence of attrition and survivor bias.

3.5.2 Missing Values

The NACC dataset is imperfect and contains many missing values for subjects which must be resolved in data pre-processing. To understand the extent of the problem, we remove features which were missing for more than 20% of participants and then visualise the percentage of participants with at least one missing data point for any of the remaining features in **Figure 8**.

It is immediately obvious that all cohorts suffer from significant missing data, even after filtering out features with the highest missing data rates. This may be the result of some features with an abnormally high percentage of missing data. The NACC backfills some fields in later visits which may explain the initial decrease. The increase may be a result of changes in data collection over time where new features are introduced for collection on initial or early visits and are therefore absent for longer participating subjects. Missing data is clearly a problem which will need to be addressed in data pre-processing.

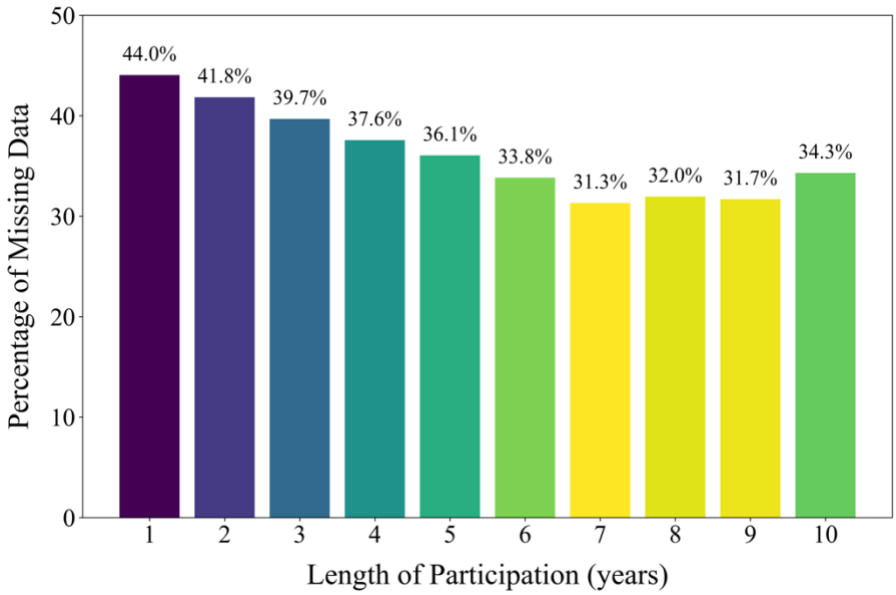

Figure 8: *Percentage of Missing Data by Length of Participation.* Shows the percentage of participants who have at least one missing data value for the features under consideration. Trends down sharply to the 7 year cohort before gradually increasing. Some NACC values such as demographics may be backfilled from subsequent visits, explaining the decrease. The later increase may be due to changes in data collection methods over time.

3.5.3 Three-year Cohort Feature Analysis

We now take a deeper dive into one cohort to gain understanding into feature patterns for medications, demographics, and some known risk factors. We emphasise that we are gaining only an illustrative insight into one cohort as it would not be feasible to explore all cohorts or features in depth, Nevertheless, identifying issues will help us to critically assess the data as a whole. We chose to assess the three-year cohort since it meets Ansart et al's (2021) minimum period criteria while maintain a balance between total sample size and progression rate.

3.5.3.1 Data representation of medications

Both categorical features, such as gender, and binary features, such as whether a medication is being taken or not, may be subject to poor frequency in data. For some binary data, the feature could be particularly uncommon, such as a treatment for a rare disease. These features may be useful for training our model on niche effects, but their effect may be difficult to validate due to a lack of statistical significance.

To understand if this is a potential issue for our data, in **Figure 9** we visualise the percentage of participants in this cohort taking a medication type.

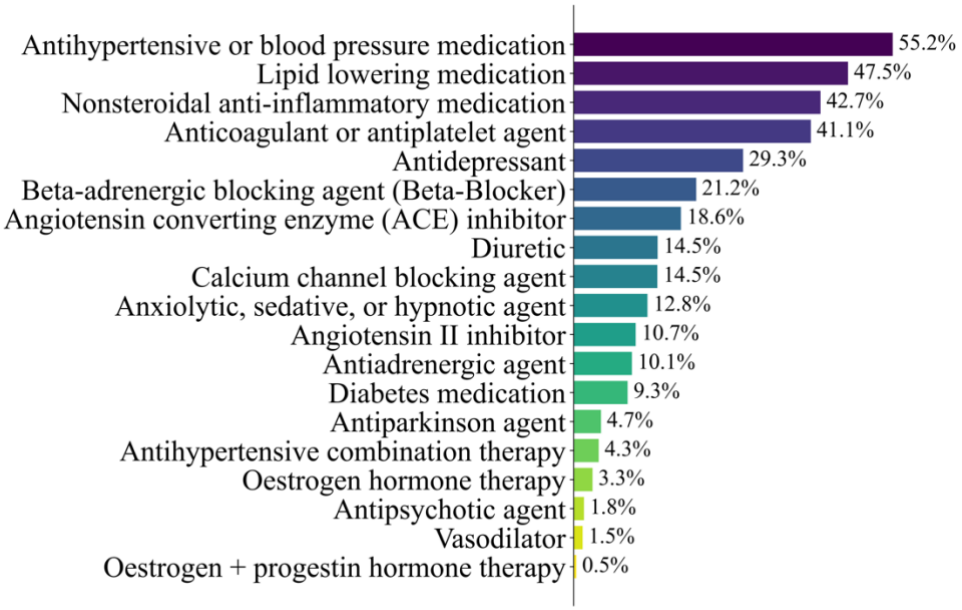

Figure 9: *Percentage of Participants Taking Each Medication.* Shows the percentage of participants taking each medication group represented by a binary feature in the 3-year cohort dataset. Around half of the medications have representation above 10% but the other medications have poor representation making inference challenging.

We observe that slightly more than half of the medications are being taken by at least 10% of study participants. Antihypertensive blood pressure medication is well represented whereas other medications of interest such as oestrogen hormone therapy and diabetes medication are poorly represented. We must exercise caution in interpreting results for poorly represented features; for example, since the impact of oestrogen hormone therapy on dementia risk is currently inconclusive (Livingston et al., 2017; Rubinstein et al., 2021; Ali et al., 2023), and only 0.46% of this cohort have this feature, any related findings are unlikely to be statistically significant.

Examining the distribution of medication across this cohort has shown that many features are imbalanced which may impact the model's ability to learn from them.

3.5.3.2 Biases in demographic features present in the data

We next assess the cohort for bias within the total dataset and patterns of progression by examining three demographic features: gender, race, and education. Relevant diagrams can be found in **Appendix B**. We observe biases in all three.

More women (34.3%) progress to dementia than men (33.4%) in this cohort, which agrees with our understanding that women are at a higher risk of dementia as they age (Ali et al., 2023). However, women represent only 48.1% of the overall data, creating a bias to male participants.

The data is heavily biased to White participants (85.8%) and, discounting severely underrepresented races, the progressed group is also proportionally biased towards White, and to a lesser extent Asian, participants who show 13.2 and 8.9 greater percentage point progression rates, respectively, than the next highest group. These disparities do not agree with known differences in race-related risks; Black or African American individuals are twice as likely as White Americans to develop dementia (Mayeda et al., 2016).

The dataset also contains a very high proportion of highly educated individuals with 35.4% attaining a doctorate level education and a very low proportion of subjects educated up to high school level (4.5%). This is likely a selection bias issue with NACC participants. We see that progression does decline as education increases apart from the "up to high school" group which is probably due to poor representation.

By examining demographic patterns in the 3-year cohort, we have seen that it is biased towards White, highly educated, and male subjects. Although we have not performed the same analysis on all cohorts, it is likely that similar issues will exist since all participants are recruited through the same processes. We state explicitly that any model trained on this data will suffer from bias and not be fit for use in a general clinical setting. Developing improved datasets is an important goal for future studies.

3.5.3.3 Patterns in risk factors present in the data

In this final EDA section, we will explore two features, Beta Blocker use and APOE4 allele count, which represent known dementia risk factors to assess if the data follows established patterns of dementia progression and what that may mean for our model. Relevant diagrams are provided in **Appendix C**.

Progression rate was 3.5 percentage points lower in the participant group who take Beta Blockers in this data set. This agrees with literature which posits that Beta Blocker use reduces the risk of AD by enhancing CSF flow to clear amyloid-beta and tau build-up (Beaman et al., 2022).

The presence of one APOE4 allele increases likelihood of progression in this cohort by over fifty percent, and presence of two alleles by one hundred percent, compared to none. Again, we see a clear agreement with the literature that an increase in e4 allele count significantly increases the likelihood of dementia progression (Livingston et al., 2017). We note, however, that those with only two alleles represent less than nine percent of the cohort and so this risk may be under considered as a model learns.

3.5.4 Summary of EDA findings

At a cohort level, EDA has revealed trade-offs between prediction period and sample size, and that the data exhibits class imbalances, age outliers and significant missing data issues. At a lower level, we have observed gender, racial and educational biases in this cohort as well as disparity in feature representation such as co-morbidities and APOE4 allele count. We can assume that similar disparities are likely in other cohorts.

These findings show that we should deal with missing data and outliers in our data pre-processing and model design and that the training data prevents the model from being clinically deployable due to severe biases.

3.6 Imbalanced Data Remediation

Classifiers trained on imbalanced data tend to have higher accuracy for majority classes, leading to a higher risk of false negatives for even small imbalances (Mazurowski et al., 2008; Fernández et al., 2018). Fernández et al (2018) describe three main strategies for mitigating data imbalance relevant to neural networks: over or under sampling of data, cost-sensitive learning and ensemble learning.

The key drawback to sampling is the introduction of noise. Our model learns from higher order relationships between patients, so removing majority cases would remove valuable data about protective relationships while duplicating minority cases could

lead to false inference from artificially overrepresented relationships. Indeed, Hassanet et al (2022) studied over seventy sampling methods and found that even the best performing techniques introduced significant errors. Therefore, given the dependence on higher order relationships, we will not use sampling.

Cost-sensitive learning is the use of a non-standard loss function in a neural network which penalizes misclassification of one class more strongly than another (Fernández et al., 2018). Adding a misclassification cost to our model for minority (positive) cases not only helps to improve accuracy by mitigating the data imbalance but also reduces the risk of classifying false negatives. Cost-sensitive learning is, therefore, a suitable technique for our model to mitigate data imbalance.

Ensemble learning is the use of multiple classifying models whose results are in some way combined to make a final classification decision (Fernández et al., 2018) and have been found effective in graph learning (Goyal et al., 2019). The underlying principle is that every model has some underlying error which can be smoothed when the results of different models are combined. To be successful, an ensemble must be composed of models which produce different errors. In our model design, the addition of the explainability module leverages ensemble learning since the models are trained together using combined loss functions but produce different errors.

Therefore, to preserve natural higher order relationships for model learning, we will remediate the imbalanced data in model design, rather than data pre-processing.

3.7 Data Pre-processing

We must ensure that the input data for our model is both appropriately structured and, to the best possible extent, without defects. We will discuss remediation of missing values, features selection, and data transformation for a hypergraph neural network.

3.7.1 Feature Selection and Missing Data

The next challenge is to select features from the total dataset provided by the NACC and to deal with missing data. The raw dataset contains 792 data columns and a row for each participant visit. A breakdown of each section can be found in **Appendix D**. The data is grouped into sections such as administrative data, medical history, and clinician diagnoses.

Since we are only using clinical and genetic features, we first remove administrative and biomarker fields, reducing the total to just over six hundred features. We then remove features which have derived equivalents. For example, although specific medications are provided as individual features, these have been collated into summary

columns by their primary effect, e.g. the use of antidepressants by a subject. Wherever a clinician diagnosis is provided alongside a patient reported condition, the clinician diagnosis is chosen. Some fields, such as the presence of certain inherited mutations, combine ‘not present’ and ‘unknown’ into the same field value and so cannot be reliably represented and are, therefore, removed. After this process, 189 clinical and genetic features remain, but many subjects still have missing values.

To deal with this missing data without sampling we must remove participants or data columns with missing values. Removing all participants with missing data results in 3,076 remaining participants. To increase the sample size, we find a balance between removing participant data and removing feature columns with missing data. We find a sweet spot for retaining features while minimizing the number of subjects to remove by keeping features where no more than 20% of participants had a missing value. Through this process we increase the sample size to 4,755 subjects with 138 features. We will conduct experiments using both datasets to analyse the trade-off between feature numbers and sample size.

3.7.2 Feature Engineering for the Hypergraph and Data Splits

Hypergraphs can be generated from any data that can be expressed as a relationship between entities. Hypergraph structure is essentially the result of hyperedge generation which may be performed explicitly where underlying data has an inherent and relevant structure to the learning task, or implicitly where it does not (Gao et al., 2022). Most features in the NACC dataset are categorical and can be modelled in a hypergraph as categories with one-hot encoding. Others, such as age and education level represent ordinal numbers – each value could be one-hot encoded, but this may lead to excessive granularity in the data which would inhibit the model from discovering higher level patterns. In these cases, edges were created by bucketing values into bins. For example, years of education was transformed into one-hot encoded education level hyperedges based on guidelines provided by the NACC.

Finally, the data was separated into random training, validation, and test sets with equal proportions of classes in each set. Given the relatively large number of samples, we are able to have relatively large test and validation groups and so chose a 60/20/20 split, respectively.

3.8 Model Design and Evaluation

To realise an explainable HGNN for dementia prediction, we use EHNN, a cutting-edge hypergraph neural network architecture created by Kim et al (2022) and augment it with an interpretable subset learning module as defined by Xu et al (2022). In this

section, we give an overview of these methods and how we have combined them. For an in-depth understanding, we direct the reader to the respective articles.

3.8.1 Equivariant Hypergraph Neural Network Transformer

At the foundation of EHNN is the novel idea to decompose undirected hypergraphs as a sequence of tensors; since a uniform hypergraph can be represented as a symmetric higher-order tensor, each k-order hypergraph can then be represented by a permutation-invariant higher-order tensor A^k (Kim et al., 2022).

Crucially, this allows the problem of creating a neural network to learn from hypergraph data to be reduced to finding a function f which is invariant and equivariant under node permutations and operates on a sequence of tensors. A neural network can thus be constructed from a sequence of equivariant linear layers as characterized by Maron et al (2019; cited by Kim et al., 2022). These layers identify equivalence classes which partition multi-index space, dictating the weight and bias parameters of the layer and are defined in equation 1 (Kim et al., 2022):

$$L_{(k)\to(l)}\left(\mathbf{A}^{(k)}\right)_{\mathbf{j}} = \mathbb{1}_{|\mathbf{j}|=l}\left(\sum_{\mathcal{J}=1}^{\min(k,l)}\sum_{\mathbf{i}}\mathbb{1}_{|\mathbf{i}\cap\mathbf{j}|=\mathcal{J}}\,\mathbf{A}_{\mathbf{i}}^{(k)}w_{\mathcal{J}} + \sum_{\mathbf{i}}\mathbf{A}_{\mathbf{i}}^{(k)}w_0 + b_l\right) \quad (1)$$

Where $w_0, w_{\mathcal{J}}, b_l$ are weight and bias. By using pairwise linear layers between input and output tensor sequences, equivariant linear layers for hypergraphs can then be composed as in equation 2 (Kim et al., 2022):

$$L_{(:K)\to(:L)}\left(\mathbf{A}^{(:K)}\right) = \left(\sum_{k\le K} L_{(k)\to(l)}\left(\mathbf{A}^{(k)}\right)\right)_{l\le L} \quad (2)$$

These layers, however, cannot yet be used in a practical model since they cannot take hypergraphs with orders exceeding (K, L) and result in the number of parameters growing at least linearly with (K, L) (Kim et al., 2022). To overcome this, Kim et al, (2022) introduce hypernetworks as a means to share trainable parameters within layers and define EHNN layers as in equation 3:

$$\begin{aligned}\mathrm{EHNN}\left(\mathbf{A}^{(:K)}\right)_{l,\mathbf{j}} = {} & 1_{|\mathbf{j}|=l}\sum_{k\le K}\sum_{\mathcal{J}=1}^{\min(k,l)}\sum_{\mathbf{i}} 1_{|\mathbf{i}\cap\mathbf{j}|=\mathcal{J}}\,\mathbf{A}_{\mathbf{i}}^{(k)}\mathcal{W}(k,l,\mathcal{J}) \\ & + 1_{|\mathbf{j}|=l}\sum_{k\le K}\sum_{\mathbf{i}}\mathbf{A}_{\mathbf{i}}^{(k)}W(k,l,0) \; + 1_{|\mathbf{j}|=l}\,\mathcal{B}(l)\end{aligned} \quad (3)$$

Where $\mathcal{W}(k,l,\mathcal{I}),\mathcal{B}(l)$ are hypernetworks inferring weights and biases. EHNN can then be realized with an attention transformer, introducing sophisticated sum pooling as in equation 4 (Kim et al., 2022):

$$\text{Attn}\left(\mathbf{A}^{(:K)}\right)_{l,\mathbf{j}} = \phi_3\left(l,\sum_{\mathcal{I}\geq 0}\phi_2\left(\mathcal{I},\sum_{h=1}^{H}\sum_{k\geq K}\sum_{\mathbf{i}}\alpha_{\mathbf{i},\mathbf{j}}^{h,\mathcal{I}}\phi_1\left(k,\mathbf{A}_{\mathbf{i}}^{(k)}\right)w_h^V\right)\right) \quad (4)$$

with $\phi_{1:3}$ as Multi-Layer Perceptron (MLP) universal approximators modelling the decomposition of the hyper-network $W_{(k,l\,\mathcal{I})}$, a trick which eliminates the need to store weights for each triplet, and $\alpha_{i,j}^{h,I}$ denoting attention coefficients calculated though scaled dot-product attention on query and key hyper-networks. The EHNN-Transformer is then given as equation 5 (Kim et al., 2022):

$$\text{EHNN-Transformer}\left(\mathbf{A}^{(:K)}\right) = \text{Attn}\left(\mathbf{A}^{(:K)}\right) + \text{MLP}\left(\mathbf{A}^{(:K)}\right) \quad (5)$$

We train the EHNN-Transformer using a binary cross-entropy loss function. To implement cost-sensitive learning, we rescale weights according to the data imbalance, increasing the weight for the minority group.

3.8.2 Interpretable Subset Learner Module

To explain the factual and counterfactual reasoning of the EHNN-Transformer, we define a module which aims to learn the subset of important edges for each node, which is defined as $\mathcal{G}'$ where $\mathcal{G}$ is the original hypergraph. Following Xu et al, (2022) $\mathcal{G}'$ should exhibit sufficiency and necessity such that prediction with $\mathcal{G}'$ should be consistent through factual reasoning and predictions with $\mathcal{G}\backslash\mathcal{G}'$ should yield opposite predictions through counterfactual reasoning.

To find $\mathcal{G}'$ we assign a random variable from the Bernoulli distribution $p_{v,e} \sim Bern(w_{v,e})$ where edge *e* is preserved for node *v* if $p_{v,e} > 0.5$ (Xu et al., 2022). The subset learner is constructed as a 2-layer MLP (g_θ) to parametrize the probability weight $(w_{v,e})$. To train the MLP, the Gumbel-max trick is used to differentiate $p_{v,e}$ from the forward pass weights $w_{v,e}$ of the MLP (Xu et al., 2022). This then yields the predictions for factual and counterfactual reasoning for each node as in equation 6 (Xu et al., 2022):

$$\hat{y}_f = f_\theta(\mathcal{G}');\ \hat{y}_{cf} = f_\theta(\mathcal{G}\backslash\mathcal{G}') \quad (6)$$

Where f_θ denotes the base model, in our case the EHNN-Transformer.

The factual loss can then be defined as in equation 7 (Xu et al., 2022):

$$\ell_f = \begin{cases} [\gamma + \hat{y}_v - \hat{y}_f]_+, & if\ y_v = 1; \\ [\gamma + \hat{y}_f - \hat{y}_v]_+, & else. \end{cases} \quad (7)$$

and the counterfactual loss as in equation 8 (Xu et al., 2022):

$$\ell_{cf} = \begin{cases} [\gamma + \hat{y}_{cf} - \hat{y}_v]_+, & if\ y_v = 1; \\ [\gamma + \hat{y}_v - \hat{y}_{cf}]_+, & else. \end{cases} \quad (8)$$

Where $[x]_+$ is the maximum of x and 0 and γ is the predefined threshold 0.5. To force g_θ to learn concise subsets, a regularization term on the weight $w_{v,e}$ is added, giving the loss function in equation 9 (Xu et al., 2022):

$$\mathcal{L}_g = \mathbb{E}_{v\sim p(\mathcal{V}_e)}\mathbb{E}_{e\sim p(\varepsilon)}[\alpha\ell_f + (1-\alpha)\ell_{cf} + \lambda_v w_{v,e}] \quad (9)$$

where α and λ_v are tunable hyperparameters controlling the balance of factual and counterfactual learning and the strength of regularization, respectively.

3.8.3 Composite Model

We then define a composite model which modifies the binary cross-entropy loss of f_θ to include factual and counterfactual loss as:

$$\mathcal{L}_f = \mathbb{E}_{v\sim p(\mathcal{V}_e)}\{BCELoss* + \lambda_m \mathbb{E}_{e\sim p(\varepsilon)}[\alpha\ell_f + (1-\alpha)\ell_{cf}]\} \quad (10)$$

where $BCELoss*$ is binary cross-entropy loss modified with cost-sensitive learning scaled to the difference between majority and minority classes and λ_m is a tunable hyperparameter which influences the strength of the factual and counterfactual loss over the base model's loss. Following Xu et al (2022), we use alternate gradient descent to train f_θ and g_θ and train f_θ with ten warmup epochs in order to improve stability. In each subsequent epoch, we first fix f_θ to extract the node prediction logits and use these to generate factual and counterfactual subsets $\mathcal{G}$ and $\mathcal{G}\backslash\mathcal{G}'$. Then, we calculate factual and counterfactual predictions by evaluating the model with an augmented incidence matrix by preserving only the edge-node relationships in the factual and counterfactual subsets, respectively. These predictions are used to calculate the loss in equation 9 and train g_θ. Next, we train f_θ fixing g_θ with the loss defined in equation 10. This cycle is repeated for each epoch to train the composite model. This process is summarised in **Figure 10**.

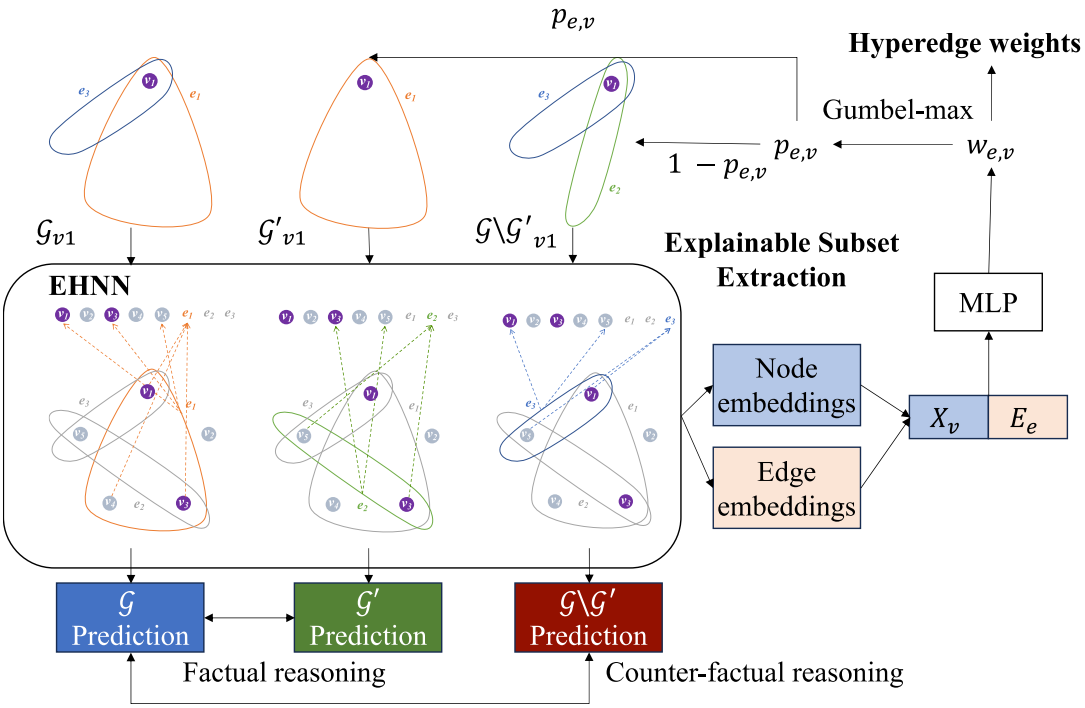

Figure 10: *Overview of the Composite EHNN and Subset Learner Model.* Adapted from Xu et al (2022). Predictions and explainable subsets are learned in alternate training cycles from the node and edge embeddings of the EHNN and factual and counter-factual predictions.

3.8.4 Model Evaluation Methods

To assess the model's performance, we need to choose a method to judge its ability to classify dementia progression correctly. Accuracy is perhaps the easiest to understand but tells us nothing about how the model is performing for different cases which can be misleading with imbalanced datasets. For example, if the data has only 20% of a positive class, by categorising all data as negative, the model can achieve an accuracy of 80% (Fernández et al., 2018). Clinicians may be more interested in the model's ability to correctly predict true positive cases or in having a low number of false positives, for example when the outcome of a positive result may be invasive follow-ups (Dubois et al., 2015). In this study we choose the F_1 score as the primary metric of model performance, the F_1 score is defined by Fernàndez et al (2018, p.52) as

> *"a weighed harmonic mean between positive predictive value and true positive rate, also known as precision and recall, respectively, ... Precision evaluates the fraction of correct classified instances among the ones classified as positive, while recall is the fraction of total positive instances correctly classified as positive"*

The F_1 score thus gives us a balanced understanding of the model's skill in identifying positive cases without over-classifying as positive. We will also consider two additional metrics: Area Under the Receiver Operating Characteristic Curve (AUC_{ROC}) and Area Under the Precision-Recall Curve (AUC_{PR}). The ROC curve is a plot of false positive rate against true positive rate and the PR curve is a plot of precision and recall, the curves show these values across all classification thresholds (Davis and Goadrich, 2006). Measuring AUC_{ROC} gives us an understanding of the model's ability to distinguish positive and negative classes whereas AUC_{PR} shows the trade-off between precision and recall. AUC_{ROC} is a popular method that is widely reported and allows us to compare against existing studies. However, in the same way as accuracy, the AUC_{ROC} can be misleading when the dataset is imbalanced (Fernández et al., 2018, p.54). Therefore, we also report the AUC_{PR} as support and visualisation for the model's performance in classifying positive cases without excessive false positives or negatives.

Together, with these scores we can assess the model's capabilities in contexts which may be desirable in different clinical contexts: overall ability to predict positive dementia progression cases (F_1 score), ability to correctly distinguishing between positive and negative cases (AUC_{ROC}), and ability to predict progression while minimising false negatives (AUC_{PR}).

3.8.5 Deterministic Results and Reproducibility

ML models often contain various sources of randomness resulting in non-deterministic results. Zhuang et al (2021) list initialization, data augmentation, data shuffling and stochastic layers as main sources of noise. Non-determinism creates two key problems: results are not reproducible and hyper-parameter or data variations cannot be fairly compared. To overcome these, we seed all random number generators used in the code and use PyTorch's deterministic flag to avoid randomness in GPU operations.

In this section we set out a specific and clinically relevant problem for our model, chose, analysed and processed the NACC dataset and defined the hypergraph neural network design. Further, we explained how we will assess the performance of our model with different measures which provide a variety of meaningful interpretations for clinicians and how we will make our results reproducible. We now use this model and data to conduct a series of experiment to answer our research questions.

Chapter 4

Experiments, Results and Analysis

Having explored the underlying data and defined our models, we now set out our contributions through experiments. We start by creating a performance baseline with only the EHNN model, answering our first research question. We then build on these with the subset learner and analyse its outputs to answer the second and third research questions. All experiments were conducted on an NVIDEA V100 GPU.

4.1 Contribution 1: Hypergraph neural networks can outperform existing methods for dementia progression prediction with clinical and genetic features *(RQ1)*.

We find that the EHNN model achieves an F_1 score of 0.73, AUC_{ROC} of 0.86, AUC_{PR} of 0.75 against a baseline of 0.34 and 79% accuracy, which outperforms non-hypergraph models in two similar studies. This section details the experiments leading to these results and their comparison to existing studies.

4.1.1 Grid search experiments

To approximate the best hyper-parameters, we first conduct a grid search varying hyper-parameter values: learning rate (LR), weight decay (WD), number of self-attention heads (SAHs), hyper-network dropout rate (HDO) and hidden layer channels (HLCs). As it would not be computationally feasible to conduct an exhaustive search across all cohorts, we search on the three-year cohort with the feature subset set selected for low number of missing values, as a trade-off between a larger number of progressed subjects and total sample size, over 500 epochs. A full table of results for both LRs is provided in **Appendix E**.

Although we see a better F_1 score with the lower LR, when we examine loss curves for the best result with each LR we find that the model overfits with the higher rate, supported by the observation that convergence is much faster when LR is lower. These loss curves can also be found in **Appendix E**. Thus, we select hyper-parameters of a LR of 0.001, 0 WD, 4 SAHs, 0,1 DO and 256 HLCs, achieving an F_1 score of 0.73. We next use these hyper-parameters to search for the best prediction period and feature set.

4.1.2 Optimal prediction period experiments

We now use the best performing hyper-parameters to search for the period which yields the highest performance. Since convergence occurs early with these hyper-parameters, we do not search beyond 200 epochs. **Table 1** shows the best performing period for each feature set. The filtered feature set, trading-off between number of subjects and number of features, performed best over a three-year period. All best performing periods have at least a three-year duration, agreeing with Ansart et al's (2021) finding that this is the minimum necessary period. For all other feature sets, we found the best time range in the five- or six-year periods, suggesting that with lower cohort sizes, a longer period is needed to observe the impact of these features. See **Appendix E** for a visualisation of each period and feature set.

These results suggest that most features play a role in the prediction, since the best performing model was trained with data from all the subsets, but also that total sample size is more important than total number of features. We also see overfitting in later time periods with smaller cohort sizes. We caveat these results on having used a set of hyper-parameters that we know to be approximately optimal for the 3-year dataset.

Feature Set	Year	VF_1	TF_1	Epoch
All selected features	5	0.72	0.71	70
All selected features (filtered to low missing values)	3	0.73	0.73	93
CDR sub-scores	5	0.69	0.68	85
Co-morbidities	6	0.58	0.57	61
Dependence related features	5	0.66	0.63	60
Family dementia and genetic features	6	0.60	0.56	58
Medications	6	0.58	0.56	60

Table 1: *Best Time Period Result for each Feature Set.* We see the best result in the 3-year time for the filtered feature set, suggesting that a sample size is more important than total features. The good performance of only CDR scores suggest these are powerful predictors.

4.1.3 Feature ablation experiments

Next, we examine performance differences when removing features from the filtered three-year dataset. Features are grouped as relating to family dementia and genetics, co-morbidities, demographics, CDR sub-scores, social isolation, health habits, behaviour and cognition characteristics, medications, dependence, and depression.

Table 2 shows the results of the feature ablation experiments on the 3-year period. We see that removing health habit features improves the validation set score but is not

supported by the test set score. For other results, we see a worse validation score and overfitting. Interestingly, we don't see a significant drop in performance when removing the CDR sub-scores, despite observing good model performance when using only them. Therefore, this supports the finding in the previous section that all feature sets contribute to model performance. This also suggests that when removing some features, the model is able to compensate.

Ablated Features	VF_1	TF_1	Epoch
Family dementia and genetic features	0.70	0.71	101
Co-morbidities	0.69	0.71	74
Demographics	0.7	0.68	46
CDR sub-scores	0.70	0.70	88
Social isolation	0.69	0.70	99
Health habits	0.73	0.71	71
Behaviour and cognition characteristics	0.70	0.70	97
Medications	0.73	0.70	152
Dependence	0.72	0.69	77
Depression	0.72	0.70	139

Table 2: *Feature Ablation Results.* Shows validation and test set F_1 scores for the filtered three-year dataset with feature subsets removed. Although ablating some subsets, e.g. Health habits, increased performance in the validation set, we see overfitting as test set results do not support them. Removing any subset – and in particular CDR scores – does not significantly reduce performance. This suggests that all feature sets contribute to model performance.

In a task to predict dementia prediction over a three-year period, EHNN achieves an AUC_{ROC} of 0.86 with an LR of 0.001 no WD, 4 SAHs, 0.1 DO and 256 HLCs. This outperforms two similar studies from Lin et al, (2018) who report a best AUC_{ROC} of 0.75 and Pang et al, who achieve an AUC_{ROC} of 0.85. Our results and these studies are compared in more detail in **section 4.4**. We achieved this result by performing a wide search across a feature sets and time periods to find the optimal tradeoff between sample size, prediction period and feature sets.

In the following section, we will include a self-explainability module and assess whether or not it is possible to maintain the same performance.

4.2 Contribution 2: We can maintain dementia progression risk prediction performance with the hypergraph neural network when including a self-explainability module. *(RQ2).*

We find that we can maintain or improve model performance when implementing a self-explainability module. We achieve an equal F_1 score of 0.73 and slightly higher AUC_{ROC} and AUC_{PR} scores of 0.87, of 0.76 respectively, compared to 0.86 and 0.75 for the EHNN only model. This section details how these scores were achieved.

To approximate the best performing composite model, we perform hyper-parameter tuning specific to the explainability module and then select the best performing parameters and compare the performance with the base model.

4.2.1 Lambda tuning experiments

The subset module and EHNN model lambda values work together to influence both the total size of the factual subset and its influence on the model's learning. **Table 3** shows results of adjusting only the EHNN λ_m hyper-parameter while keeping others fixed. We find the best value to be 0.1. When the value is too low, the model fails to incorporate the factual and counterfactual loss into its learning, and when the value is too high, the model is influenced too much by this loss in earlier epochs.

EHNN λ_m	VF_1	TF_1	AUC_{ROC}	AUC_{PR}
10	0.71	0.73	0.86	0.74
1.5	0.72	0.73	0.86	0.74
1	0.72	0.71	0.86	0.75
0.5	0.73	0.74	0.86	0.75
0.1	0.73	0.74	0.86	0.76
0.01	0.72	0.71	0.86	0.74

Table 3: *EHNN Lambda Tuning.* Shows performance with different values for EHNN model lambda. Subset learner lambda decay rate and alpha are fixed. We see the best performance across all metrics with a value of 0.1. If the value is too high, the effect on model learning is too much, and if it is too low, too little.

The subset learner lambda value determines regularization strength for the generated subset: higher values will lead to smaller subsets and vice versa. It's desirable for this impact to be high in early training epochs, allowing the model to quickly discard irrelevant information; however, if left constant, it will lose too much relevant information. To avoid this, we first set an initially high value which encourages stricter

subsets then decay this value rapidly. Results of this tuning are displayed in **Table 4**. This process revealed that an initial subset learner lambda value of $1E^{-8}$ with a decay of 14% per epoch achieved the best results and slightly outperforms the EHNN model on AUC_{ROC} and AUC_{PR} by 0.01.

λ_sDR	VF_1	TF_1	AUC_{ROC}	AUC_{PR}
12%	0.72	0.74	0.86	0.76
14%	0.73	0.74	0.87	0.76
16%	0.72	0.74	0.86	0.76
18%	0.72	0.72	0.86	0.73
20%	0.72	0.71	0.86	0.74
22%	0.73	0.71	0.86	0.76
24%	0.72	0.74	0.86	0.76
26%	0.72	0.74	0.87	0.76

Table 4: *Subset Learner Lambda Decay Rate Tuning.* We see best performance at 14% decay, suggesting less strict subsets yield better performance. Moreover, this performance slightly exceeds the base EHNN mode performance on AUC_{ROC} and AUC_{PR} by 0.01.

4.2.2 Alpha tuning experiments

The alpha value determines the weight balance of factual and counterfactual loss in model learning (Xu et al., 2022); a higher alpha value places more weight on factual reasoning and less on counter-factual reasoning. **Table 5** shows that an equal balance of factual and counter-factual learning yields the best results.

α	VF_1	TF_1	AUC_{ROC}	AUC_{PR}
0	0.72	0.72	0.86	0.75
0.25	0.72	0.73	0.86	0.75
0.5	0.73	0.74	0.87	0.76
0.75	0.72	0.73	0.86	0.75
1	0.72	0.72	0.85	0.73

Table 5: *Alpha Tuning.* Alpha controls the balance of factual and counter-factual learning – a value of 1 allows only for factual learning and vice versa. We see the best results when equally balancing factual and counterfactual results, although differences are minor.

These experiments have shown that including a self-explainability module not only maintains but also improves module performance of the model for AUC_{ROC} and AUC_{PR} compared to the stand-alone module.

We next examine the feasibility of using the outputs of the explainability module to interpret its predictions.

4.3 Contribution 3: We can assess the explainability-augmented hypergraph neural network's limitations by examining the self-explainability module's explanations for its dementia risk classifications. *(RQ3)*

Having implemented a self-explainability subset learner module, we now show that its output can be used to assess the model's strengths and weaknesses in reasoning.

4.3.1 Assessing hyperedge importance

The module learns by assigning weights to each node-hyperedge connection. We can rank these scores for each subject as a form of explanation for its prediction. To understand underlying patterns in these subsets, we aggregate and rank these weights for correctly predicted positive cases across all patients.

Since hypergraphs consider complex relationships between multiple features, it is not sufficient to look at only the average weights. A weight which is significant for one patient with a specific subset of co-morbidities may be less significant for a patient with a different set, so large hyperedges are likely to have a wider range of associated weight values. To demonstrate, **Figure 11** shows a plot of the range against the total number of hyperedge-node instance counts for each hyperedge, considering only correctly predicted positive cases. This shows that large hyperedges are more likely to have a wider range of associated weights. Therefore, to identify edges with a relatively higher impact, we create a weighted score for each hyperedge e_i as a sum of the min-max normalized average weight and the percentage of weights in the hyperedge which are greater than the overall average of all weights in all hyperedges in equation 11:

$$score_{e_i} = \frac{w_{e_i} - w_{e_{min}}}{w_{e_{max}} - w_{e_{max}}} + \frac{|e_i > \bar{E}|}{|e_i|} \quad (11)$$

Normalization allows us to express both parts of the score with a zero to one scaled number, so we can use this score to rank hyperedges. We assess the results of this ranking in the next section.

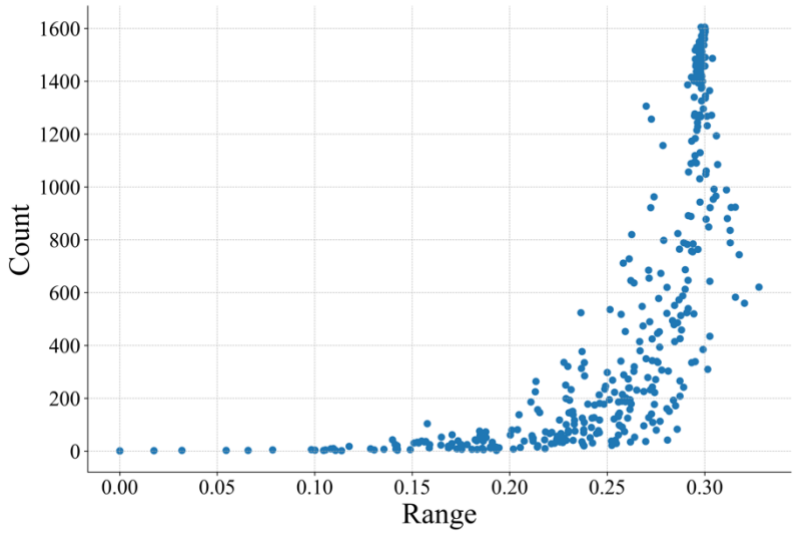

Figure 11: *Scatter Plot of Hyperedge Weight Range against Count of Relationships.* Shows how range varies by the count of nodes for each edge. Hyperedges with more nodes have a larger range of hyperedge scores since each patient has a unique risk relationship.

4.3.2 Hyperedge ranking analysis.

We now use the hyperedge scores to gain understanding into the models reasoning. There are 396 hyperedges in the hypergraph, so we pick the top scoring hyperedges for analysis. The distribution is skewed: out of 396 hyperedges, only 55 (14%) score above one. We examine the top thirteen hyperedges, which scored over 1.5. A histogram of edge scores can be found in **Appendix E**.

Table 6 shows these hyperedges, ranked by score. The most notable result in the table is that the edge encoding zero APOE4 alleles ranked 3rd despite the edge encoding presence of a dominantly inherited AD gene ranking 13th. While this seems contradictory, this result may be a result of how the model learns. In this analysis we are only able to assess the overall hyperedge scores; however, the model learns through higher-order relationships between hyperedges. Since the presence of APOE4 is likely a strong signal in isolation, the model may be strengthening the signal of the absence of APOE4 in relationship to other factors. To support this, we examine the strength of the APOE4 allele count hyperedge scores for each number of APOE4 alleles in **Figure 12**. This shows comparative histograms of score distributions for each hyperedge smoothed with kernel density estimation for subjects by their number of APOE4 alleles. The hyperedge representing no APOE4 alleles has a significantly wider range and higher average of edge scores than those representing one or two alleles present.

This suggests that having no APOE4 allele is a signal to the model to focus on other relationships for classification. However, to conclusively answer this question, it would be necessary to develop a method to explain how edge scores relate to each other for different participants, which is outside of the scope of this study.

Other scores are more in line with expectations – we see four examples of CDR scores in the top-ranking edges. There are also behavioural and cognitive signals, such as a lack of engagement in normal day to day activities, and links to frontotemporal lobar degeneration genealogy. We also see a relatively younger age range, which is likely receiving a high edge score for the same reason as the no APOE4 allele group.

	Code	Meaning	Score	Count
1	_NACCFAM_0	No report of a first-degree family member with cognitive impairment	2.00	560
2	_NACCFTLD_1	In this family, is there evidence for a frontotemporal lobar degeneration mutation?	1.65	7
3	_E4_0	No E4 APOE alleles	1.65	621
4	_JUDGMENT_3	Judgment and problem-solving: Severe impairment (CDR Score 3)	1.61	1
5	_JUDGMENT_2	Judgment and problem-solving: Moderate impairment (CDR Score 2)	1.61	11
6	_COMMUN_2	Community affairs: Moderate impairment (CDR Score 2)	1.59	3
7	_ORIENT_2	Orientation: Moderate impairment (CDR Score 2)	1.57	3
8	_NACCCOGF_7	Fluctuating cognition noticed as first sign of impairment	1.56	1
9	_NACCFTDM_1	Subject has a hereditary frontotemporal lobar degeneration mutation	1.56	7
10	_AGE_50_55	Age range 50 to 55	1.48	41
11	_STOVE_8	In the past 4 weeks, subject did not use stove	1.48	31
12	_SHOPPING_8	In the past 4 weeks, subject did not engage in shopping activities	1.47	53
13	_NACCADMU_1	Subject has dominantly inherited AD gene	1.42	2

Table 6: *Top 13 Ranking Hyperedge Scores Learned by the Subset Learner*. We see several expected results e.g. high CDR scores and signals of cognitive decline, but we also unexpected results such as no e4 APOE alleles. These may indicate that the model has scored these highly based on their relationship with other hyperedges.

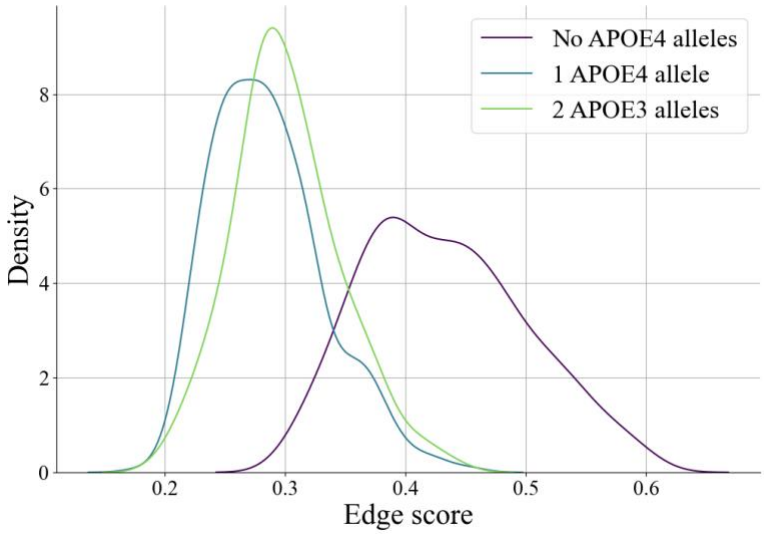

Figure 12: *Kernel Density Estimation Smoothed Histogram of Hyperedge Score Distribution for Different APOE4 Groups in the Correctly Classified Positive Class.* We observe that the hyperedge representing no APOE4 alleles had higher edge scores than the hyperedges representing one or two alleles, suggesting that this is a signal to focus on other factors.

Analysis of aggregate hyperedge scores has shown that the explainability module can provide insight into high level and complex patterns of the model's decision-making process for classifying dementia prediction. However, additional work is required to understand the precise and multi-faceted causal relationships learned by the model.

4.4 Discussion

This EHNN model alone achieved a validation set F_1 score of 0.73 supported by an equal test set F_1 score. The model's accuracy was 79%. The AUC_{ROC} of 0.86 indicates that the model has good skill in distinguishing positive and negative classes. We achieve an AUC_{PR} of 0.75. In the context of a PR curve, a no skill model on our dataset would have an AUC_{PR} of 0.34 (equal to the proportion of positive classes), so 0.75 demonstrates a good skill in classifying the positive case.

After incorporating the factual and counter-factual self-explainability module we achieve and equal F_1 score and slightly improve AUC_{ROC} abd AUC_{PR} compared to the EHNN model alone, showing that explainability does not come at the cost of performance for this model. **Figure 13** shows the AUC_{ROC} of the best-tuned composite subset and EHNN model overlayed with the best tuned EHNN-only model, and **Figure 14** shows the same for AUC_{PR}. The composite model slightly improves performance

for both AUC_{ROC} and AUC_{PR}. This suggests that incorporating factual and counterfactual learning improves the model's ability to distinguish between classes and in reducing the number of false positives and false negatives, a desirable characteristic of a model in a clinical setting. For both models, the area under the left side of the AUC_{PR} curve is higher than the right, showing that both models are better at reducing false positives than false negatives, but the composite model has a slight advantage in reducing false negatives. This information is useful for clinicians who have context on whether a false positive or negative would be more harmful for a patient.

To contextualize these results in the literature, we compare them against two recent studies which have similar design and objectives. Lin et al, (2018) performed experiments with SVM, Logistic Regression and Random Forest classification models to predict dementia progression in a 4-year period using 348 clinical, non-invasive features on the NACC dataset. They report a best AUC_{ROC} of 0.75 with a SVM model but do not report F_1 or AUC_{PR} scores. Similarly, Pang et al, (2023) conducted experiments with SVM, LR and RF models. This study had a greater focus on progress to AD from MCI and MCI from normal cognition rather than general dementia progression but attempt to do so based only on clinical features. They report an AUC_{ROC} of 0.85 in predicting progression to AD from MCI in a 2-year period and 0.8 over a 3-year period with a RF classifier. Again, no AUC_{PR} scores were reported. Our model outperforms both results in AUC_{ROC}, particularly over longer periods, by leveraging higher order relationships in the data.

We have demonstrated that a hyperedge neural network can out-perform existing models for dementia detection. Further, we have done so in a well-defined and reproducible experiment context. This supports our hypothesis that hypergraph neural networks can yield better results by leveraging complex interrelationships between features.

We also implemented a self-explainability sub-module which not only maintained model performance but also improved its capabilities in distinguishing between classes and in reducing false negatives, two clinically important tasks. We then applied rudimentary techniques on the aggregate output of the model's reasoning which shows promising insight into the model's reasoning. We found some evidence that the model is encoding complicated signals relating to higher order relationships in hyperedge weights. However, we are limited in a deeper examination of the model's reasoning as we can only observe patterns across individual hyperedge scores. Further, the model's overall performance is good, but not excellent, and we observed many issues in the EDA, so we can expect a certain amount of noise and bias to mislead this analysis. Therefore, there we suggest that this is a promising area for future research alongside improvements in data collection and further advances in model architecture.

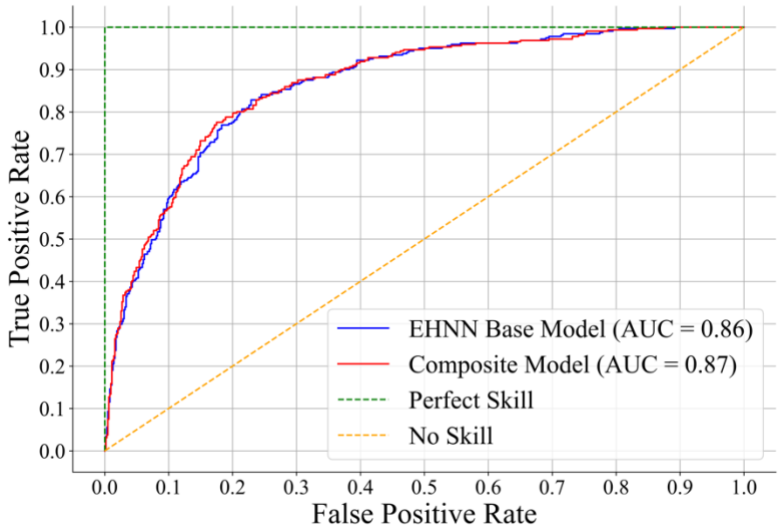

Figure 13: *Receiver Operating Characteristic Curves for Composite and EHNN-only models.* Shows that the composite, explainable model achieved very similar performance in distinguishing classes as the base model. The steep incline, followed by a gradual increase of the curve shows the model identifies a low number of FPs and a high number of TPs.

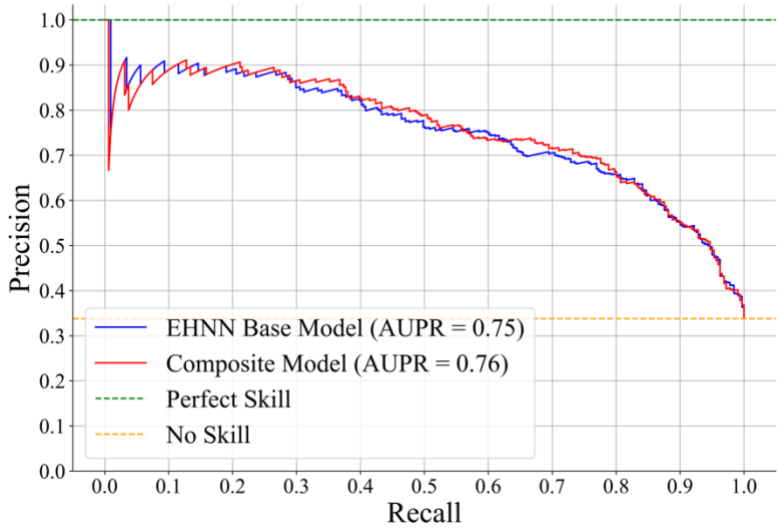

Figure 14: *Precision Recall Curves for Composite and EHNN-only models.* Shows that the composite model is slightly better at classifying the positive case than the base model. Precision (TP / TP + FP) decreases rapidly only after recall (TP / TP + FN) exceeds 0.8. Since this curve is has a larger area left of center, the models are slightly better at reducing false positives than false negatives.

Chapter 5

Conclusion and Future Work

Our study has provided the following contributions:

1. **Hypergraph neural networks can outperform existing methods for dementia progression prediction with clinical and genetic features.**
2. **Hypergraph neural network performance can be maintained when implementing self-explainability.**
3. **We can use explainability to assess the model's reasoning.**

We achieved these by training a cutting edge-model on a refined dataset and applying a self-explainability module to analyse and evaluate the model's reasoning.

The dataset included only clinical, low-cost and non-invasive features. Achieving good, clinically relevant results with this data has significant clinical implications. First, such a tool can provide clinicians with a means to pre-screen patients for dementia risk before using invasive and expensive methods, reducing the ethical decision making burden (Van Der Schaar et al., 2022). Further, dementia prevalence is set to increase in low- and middle-income countries as a result of increasing life expectancy (Farina et al, 2023) where more expensive methods may not be feasible. The WHO is aiming for 75% of such countries to have a national dementia plan in place by 2025 (Beuer et al, 2022). Such a tool trained for specific populations can support this target by alleviating the burden of diagnosis costs where they are most sensitive. Further, as we showed that this tool may discover complex patterns, with population specific training, it may be able to discover such patterns of risk specific to a developing nation population or sub-population.

Explainability is crucial for clinician trust and can help to uncover previously unknown patterns of risk (Rajpurkar et al., 2022). This study has shown that explainability is not only possible with a dementia prediction hypergraph neural network, but also that it can improve model performance. We found patterns which agree with existing literature – such as that CDR scores alone are a good predictor for dementia progress (Kim et al, 2017), and results which seem at first to disagree with well-established risk factors: that APOE4 allele count (Livingston et al, 2017) and family history of dementia (Wolters et al, 2017) are strong predictors of dementia progression. However, further examination showed that these surprising results are most likely the result of complex patterns which the model uncovers in order to diagnose individuals

who do not exhibit these characteristics which the techniques used in this study could not fully uncover.

Indeed, the model's overall performance and the representation of features in the underlying data set show that it would be premature to draw concrete conclusions from its reasoning or to recommend deploying the model in a clinical setting. The techniques used to assess the influence of edge weights are rudimentary as we do not currently have a method which fully leverages the higher order relationships learned in the hyperedge weights. For example, although Terrosu (2022) found a relationship between atrial fibrillation and dementia, the specific mechanisms are unknown and likely multifactorial involving several co-morbidities. A module which can express these relationships would provide significant advancements in our understanding of this relationship. Further, the study is primarily limited by the data set used to train the model; it has several biases, particularly in race and education level and poor representation of many features.

Therefore, future work to progress towards a more performant model includes:

- A concerted data collection effort to produce a population-level collection of data with good feature representation and without biases.
- Exploration of self-explainability techniques for self-attention transformers which can expose learned higher order relationships for hypergraphs.

Two recent and promising studies may be adapted to develop more advanced global explanations for hypergraphs. Azzolin et al (2023) have developed an ML technique called GLGExplainer for Graph Neural Networks which uses aggregate local explanations, such as those produced in our combined model, to learn concepts as prototypical features of explanations and have demonstrated its efficacy with hospital interaction networks. This technique could be modified to work with hypergraph outputs to uncover important concepts such as specific interactions between comorbidities. Maleki et al (2023) have also developed a technique more specific to Hypergraph Neural Networks called HyperEx which aims to learn node-hyperedge importance scores by generating explanatory sub-hypergraphs. This technique is a more sophisticated approach to the method used in this paper and may provide more robust results; however, it may still suffer in explaining more complex multi-edge relationships. Nevertheless, the development of new techniques for graph and hypergraph explanation indicate that our model may be further improved so that it can be confidentially and valuably deployed in a real-world clinical setting.

In summary, although limited, this experimental project has shown the potential of hypergraph neural networks to predict dementia progression and reveal underlying patterns in a clinically relevant manner.

Bibliography

Ali, N., Sohail, R., Jaffer, S.R., Siddique, S., Kaya, B., Atowoju, I., Imran, A., Wright, W., Pamulapati, S., Choudhry, F., Akbar, A. and Khawaja, U.A., 2023. The Role of Estrogen Therapy as a Protective Factor for Alzheimer's Disease and Dementia in Postmenopausal Women: A Comprehensive Review of the Literature. *Cureus* [Online] 15(8). Available from: https://doi.org/10.7759/cureus.43053 [Accessed 19 November 2023].

Alzheimer's Society, n.d. *Genetic testing kits and dementia* [Online]. Available from: https://www.alzheimers.org.uk/about-dementia/risk-factors-and-prevention/genetic-testing-kits [Accessed 18 November 2023].

Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., Couronné, R., Faouzi, J., Koval, I., Louis, M., Thibeau-Sutre, E., Wen, J., Wild, A., Burgos, N., Dormont, D., Colliot, O. and Durrleman, S., 2021. Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. *Medical Image Analysis* [Online], 67. Available from:https://doi.org/10.1016/j.media.2020.101848.

Aviles-Rivero, A.I., Runkel, C., Papadakis, N., Kourtzi, Z. and Schönlieb, C.-B., 2022. *Multi-Modal Hypergraph Diffusion Network with Dual Prior for Alzheimer Classification* [Online]. Available from: http://arxiv.org/abs/2204.02399 [Accessed 11 March 2023].

Azzolin, S., Longa, A., Barbiero, P., Liò, P., Passerini, A., 2023, Global Explainability of GNNs via Logic Combination of Learned Concepts, *The Eleventh International Conference on Learning Representations*, 1 - 5 May 2023, Kigali, Rwanda. Appleton, WI: ICLR, pp 1 - 19

Bai, S., Zhang, F. and Torr, P.H.S., 2021. Hypergraph Convolution and Hypergraph Attention *Pattern Recognition* [Online], 110. Available from:

https://doi.org/10.1016/j.patcog.2020.107637 [Accessed 18 March 2023].

Barnes, D.E. and Lee, S.J., 2011. Predicting Alzheimer's risk: why and how? *Alzheimer's Research & Therapy*, 3(6), pp.33-35.

Basit, M.S., Khan, A., Farooq, O., Khan, Y.U. and Shameem, M., 2022. Handling Imbalanced and Overlapped Medical Datasets: A Comparative Study, *5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, 26-27 November 2022, Aligarh. Aligarh India: IEEE, pp.1–7.

Battineni, G., Chintalapudi, N., Hossain, M.A., Losco, G., Ruocco, C., Sagaro, G.G., Traini, E., Nittari, G. and Amenta, F., 2022. Artificial Intelligence Models in the Diagnosis of Adult-Onset Dementia Disorders: A Review. *Bioengineering*, 9(8), p.370-384.

Beaman, E.E., Bonde, A.N., Larsen, S.M.U., Ozenne, B., Lohela, T.J., Nedergaard, M., Gíslason, G.H., Knudsen, G.M. and Holst, S.C., 2022. Blood–brain barrier permeable β-blockers linked to lower risk of Alzheimer's disease in hypertension. *Brain*, 146(3), pp.1141–1151.

Breuer, E., Comas-Herrera, A., Freeman, E., Albanese, E., Alladi, S., Amour, R., Evans-Lacko, S., Ferri, C.P., Govia, I., García, C.I.A., Knapp, M., Lefevre, M., López-Ortega, M., Lund, C., Musyimi, C., Ndetei, D., Oliveira, D., Palmer, T., Pattabiraman, M., Sani, T.P., Taylor, D., Taylor, E., Theresia, I., Thomas, P.T., Turana, Y., Weidner, W., Schneider, M., 2022. Beyond the project: Building a strategic theory of change to address dementia care, treatment and support gaps across seven middle-income countries. *Dementia*, 21(1), pp.114-135.

Burns, A. and Levy, R., eds, 1994. *Dementia*. Boston, MA: Springer US.

Chandramouli, S., Dutt, S. and Das, A.K., 2018. *Machine Learning* [Online]. Pearson Education India. Available from: https://learning.oreilly.com/library/view/machine-learning/9789389588132/ [Accessed 28 August 2023].

Chien, E., Pan, C., Peng, J. and Milenkovic, O., 2021. You are AllSet: A Multiset Function Framework for Hypergraph Neural Networks. *The Tenth International Conference on Learning Representations*, 25-29 April 2022, Virtual, [Online]. Available from: https://openreview.net/forum?id=hpBTIv2uy_E [Accessed 15 April 2023].

Davis, J. and Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 25-29 June 2006, Pittsburgh, Pennsylvania. New York: ACM Press, pp.233–240.

Dubois, B., Padovani, A., Scheltens, P., Rossi, A. and Dell'Agnello, G., 2015. Timely Diagnosis for Alzheimer's Disease: A Literature Review on Benefits and Challenges. *Journal of Alzheimer's Disease*, 49(3), pp.617–631.

European Commission, 2021. *Artificial Intelligence in healthcare: Applications, risks, and ethical and societal impacts* [Online]. (PE 729.512). European Commission, p.69. Available from: https://digital-strategy.ec.europa.eu/en/library/artificial-intelligence-healthcare-report

Farina, N., Jacobs, R., Turana, Y., Fitri, F.I., Schneider, M., Theresia, I., Docrat, S., Sani, T.P., Augustina, L., Albanese, E., Comas-Herrera, A., Du Toit, P., Ferri, C.P., Govia, I., Ibnidris, A., Knapp, M., Banerjee, S., 2023. Comprehensive measurement of the prevalence of dementia in low- and middle-income countries: STRiDE methodology and its application in Indonesia and South Africa. *BJPsych Open*. 9(4): e102.

Fawns-Ritchie, C. and Deary, I.J., 2020. Reliability and validity of the UK Biobank cognitive tests. *PLoS ONE* [Online], 15(4), p.e0231627.
Available from: https://doi.org/10.1371/journal.pone.0231627

Feng, Y., You, H., Zhang, Z., Ji, R. and Gao, Y., 2019. Hypergraph Neural Networks, *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 27 January - 1 February 2019, Honolulu, Hawaii. Washington DC: AAAI, pp. 3558-3565

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B. and Herrera, F., 2018. *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing.

Fitzhugh, M.C. and Pa, J., 2022. Sex differences in the association between hearing impairment and brain atrophy: findings from the UK Biobank. *Alzheimer's & Dementia* [Online], 18(S6). Available from: https://doi.org/10.1002/alz.061963 [Accessed 18 March 2023].

Ford, E., Milne, R. and Curlewis, K., 2023. Ethical issues when using digital biomarkers and artificial intelligence for the early detection of dementia. *WIREs Data Mining and Knowledge Discovery* [Online], p.e1492. Available from: https://doi.org/10.1002/widm.1492.

Gao, Y., Zhang, Z., Lin, H., Zhao, X., Du, S. and Zou, C., 2022. Hypergraph Learning: Methods and Practices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5), pp.2548–2566.

Gerritsen, A.A.J., Bakker, C., Verhey, F.R.J., De Vugt, M.E., Melis, R.J.F., Koopmans, R.T.C.M., Oosterveld, S.M., Kessels, R.P., Olde Rikkert, M.G., Hamel, R., Ramakers, I.H., Aalten, P., Sistermans, N., Smits, L.L., Pijnenburg, Y.A. and Van Der Flier, W.M., 2016. Prevalence of Comorbidity in Patients With Young-Onset Alzheimer Disease Compared With Late-Onset: A Comparative Cohort Study. *Journal of the American Medical Directors Association*, 17(4), pp.318–323.

Goyal, P., Huang, D., Chhetri, S.R., Canedo, A., Shree, J. and Patterson, E., 2019. *ASONAM 2020 - The 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 7-10 December 2020, Virtual. New York: ACM Press, pp. 24-31

Hassanat, A.B., Tarawneh, A.S., Altarawneh, G.A. and Almuhaimeed, A., 2022. *Stop Oversampling for Class Imbalance Learning: A Critical Review*, IEEE Access, 10, pp. 47643-47660

Hatherley, J., Sparrow, R. and Howard, M., 2022. The Virtues of Interpretable Medical Artificial Intelligence. *Cambridge Quarterly of Healthcare Ethics*, Cambridge University Press, pp. 1–10

Hobson, P., 2019. Enabling People with Dementia: Understanding and Implementing Person-Centred Care, Cham: Springer International Publishing.

Jack, C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeberlein, S.B., Holtzman, D.M., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J.L., Montine, T., Phelps, C., Rankin, K.P., Rowe, C.C., Scheltens, P., Siemers, E., Snyder, H.M. and Sperling, R., 2018. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 14(4), pp.535–562.

Joshi, A.V., 2023. *Machine Learning and Artificial Intelligence*, Cham: Springer International Publishing.

Kakkad, J., Jannu, J., Sharma, K., Aggarwal, C. and Medya, S., 2023. *A Survey on Explainability of Graph Neural Networks* [Online]. arXiv. Available from: http://arxiv.org/abs/2306.01958 [Accessed 7 October 2023].

Kim, J., Oh, S., Cho, S. and Hong, S., 2022. Equivariant Hypergraph Neural Networks. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) *Computer Vision – ECCV 2022. Lecture Notes in Computer Science*, 13681. Cham: Springer, pp. 86-103.

Kim, J.W., Byun, M.S., Sohn, B.K., Yi, D., Seo, E.H., Choe, Y.M., Kim, S.G., Choi, H.J., Lee, J., Chee, I.S., Woo, J.I., & Lee, D.Y., 2017. Clinical Dementia Rating Orientation Score as an Excellent Predictor of the Progression to Alzheimer's Disease in Mild Cognitive Impairment. *Psychiatry Investigation*, 14, pp. 420 - 426.

Kumar, S., Oh, I., Schindler, S., Lai, A.M., Payne, P.R.O. and Gupta, A., 2021. Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review. *JAMIA Open*, 4(3), pp. 1- 10

Li, M., Zhang, Y., Li, X., Cai, L. and Yin, B., 2022. Multi-view hypergraph neural networks for student academic performance prediction. *Engineering Applications of Artificial Intelligence*, 114, pp.105174 - 105185.

Lin, M., Gong, P., Yang, T., Ye, J., Albin, R.L. and Dodge, H.H., 2018. Big Data Analytical Approaches to the NACC Dataset: Aiding Preclinical Trial Enrichment. *Alzheimer disease and associated disorders*, 32(1), pp.18–27.

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S.G., Huntley, J., Ames, D., Ballard, C., Banerjee, S., Burns, A., Cohen-Mansfield, J., Cooper, C., Fox, N., Gitlin, L.N., Howard, R., Kales, H.C., Larson, E.B., Ritchie, K., Rockwood, K., Sampson, E.L., Samus, Q., Schneider, L.S., Selbæk, G., Teri, L. and Mukadam, N., 2017. Dementia prevention, intervention, and care. *The Lancet*, 390(10113), pp.2673–2734.

MacInnes, J., 2020. Exploratory Data Analysis. *SAGE Research Methods Foundations* [Online], London : SAGE Publications Ltd.

Available from: https://doi.org/10.4135/9781526421036889602 [Accessed 27 August 2023].

Maleki, S., Hajiramezalani, E., Scalia, G., Biancalani, T., Chuang, K. V., 2023. Learning to Explain Hypergraph Neural Networks. *2nd Annual TAG in Machine Learning*, 28 July 2023, Honolulu, Hawaii. Cham: Springer, pp. 1-6

Marcos, A., Gil, P., Barabash, A., Rodriguez, R., Encinas, M., Fernández, C. and Cabranes, J.A., 2006. Neuropsychological Markers of Progression From Mild Cognitive Impairment to Alzheimer's Disease. *American Journal of Alzheimer's Disease & Other Dementias*, 21(3), pp.189–196.

Mayeda, E.R., Glymour, M.M., Quesenberry, C.P. and Whitmer, R.A., 2016. Inequalities in dementia incidence between six racial and ethnic groups over 14 years. *Alzheimer's & Dementia*, 12(3), pp.216–224.

Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A. and Tourassi, G.D., 2008. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2), pp.427–436

Morris, J.C., 1993. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*, 43, pp. 2412–2414.

Mintun, M.A., Lo, A.C., Duggan Evans, C., Wessels, A.M., Ardayfio, P.A., Andersen, S.W., Shcherbinin, S., Sparks, J., Sims, J.R., Brys, M., Apostolova, L.G., Salloway, S.P. and Skovronsky, D.M., 2021. Donanemab in Early Alzheimer's Disease. *The New England Journal of Medicine*, 384(18), pp.1691–1704.

NACC, 2015. *Form B4: CDR® Dementia Staging Instrument* [Online]. Available from: https://files.alz.washington.edu/documentation/uds3-ivp-b4.pdf [Accessed 27 August 2023].

NHS, 2018. *Alzheimer's disease*. *nhs.uk* [Online]. Available from: https://www.nhs.uk/conditions/alzheimers-disease/ [Accessed 20 March 2023].

Panesar, A., 2020. Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes. Berkley, CA: Apress.

Pang, Y., Kukull, W., Sano, M., Albin, R.L., Shen, C., Zhou, J. and Dodge, H.H., 2023. Predicting Progression from Normal to MCI and from MCI to AD Using Clinical Variables in the National Alzheimer's Coordinating Center Uniform Data Set Version 3: Application of Machine Learning Models and a Probability Calculator. *The Journal of Prevention of Alzheimer's Disease* 10(2), pp.301–313.

Pelegrini, L.N.C., Mota, G.M.P., Ramos, C.F., Jesus, E. and Vale, F.A.C., 2019. Diagnosing dementia and cognitive dysfunction in the elderly in primary health care: A systematic review. *Dementia & Neuropsychologia* 13(2), pp.144–153.

Perneczky, R., Wagenpfeil, S., Komossa, K., Grimmer, T., Diehl, J. and Kurz, A., 2006. Mapping Scores onto Stages: Mini-Mental State Examination and Clinical Dementia Rating. *The American Journal of Geriatric Psychiatry*, 14(2), pp.139–144.

Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E.J., 2022. AI in health and medicine. *Nature Medicine*, 28(1), pp.31–38.

Rowe, T.W., Katzourou, I.K., Stevenson-Hoare, J.O., Bracher-Smith, M.R., Ivanov, D.K. and Escott-Price, V., 2021. Machine learning for the life-time risk prediction of Alzheimer's disease: a systematic review. *Brain Communications* [Online], 3(4), fcab246. Available from: https://doi.org/10.1093/braincomms/fcab246.

Rubinstein, Z.B., Buckley, R.F., Scott, M.R., Manning, L.K., Mayblyum, D.V., Thibault, E.G., Jacobs, H.I.L., Farrell, M.E., Properzi, M.J., Rabin, J.S., Chhatwal, J.P., Lois, C., Rentz, D., Price, J.C., Schultz, A.P., Sperling, R.A. and Johnson, K.A., 2021. Self-reported history of estrogen hormone therapy differentiates rates of amyloid accumulation (PiB-PET) relative to males: Findings from the Harvard Aging Brain Study. *Alzheimer's & Dementia* [Online], 17(S4), p.e056069. Available from: https://doi.org/10.1002/alz.056069.

Saito, Y. and Murayama, S., 2007. Neuropathology of mild cognitive impairment. *Neuropathology*, 27(6), pp.578–584.

Shao, W., Peng, Y., Zu, C., Wang, M. and Zhang, D., 2020. Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease. *Computerized Medical Imaging and Graphics* [Online], 80, p.101663. Available from: https://doi.org/10.1016/j.compmedimag.2019.101663.

Shao, W., Xiang, S., Zhang, Z., Huang, K. and Zhang, J., 2021. Hyper-graph based sparse canonical correlation analysis for the diagnosis of Alzheimer's disease from multi-dimensional genomic data. *Methods*, 189, pp.86–94.

Tahami Monfared, A.A., Byrnes, M.J., White, L.A. and Zhang, Q., 2022. Alzheimer's Disease: Epidemiology and Clinical Progression. *Neurology and Therapy*, 11(2), pp.553–569.

Terrosu, P., 2022. Association between heart and dementia… keep an eye on the left atrium. *European Heart Journal Supplements*, 24(SI), pp.I186–I189.

Torres, L., Blevins, A.S., Bassett, D.S. and Eliassi-Rad, T., 2020. The why, how, and when of representations for complex systems, *SIAM Review*, 63(3), pp. 435-485

Van Der Schaar, J., Visser, L.N.C., Bouwman, F.H., Ket, J.C.F., Scheltens, P., Bredenoord, A.L. and Van Der Flier, W.M., 2022. Considerations regarding a diagnosis of Alzheimer's disease before dementia: a systematic review. *Alzheimer's Research & Therapy*, 14(1), pp.31-43.

Wang, X., Xin, J., Wang, Zhongyang, Li, C. and Wang, Zhiqiong, 2022. An Evolving Hypergraph Convolutional Network for the Diagnosis of Alzheimer's Disease. *Diagnostics*, 12(11), p.2632-2643.

Wolters, F.J., van der Lee, S.J., Koudstaal, P.J., van Duijn, C.M., Hofman, A., Ikram, M.K., Vernooij, M.W., Ikram, M.A., 2017. Parental family history of dementia in relation to subclinical brain disease and dementia risk. *Neurology*, 88(17) pp. 1642-1649.

WHO, 2023. *Dementia* [Online]. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia [Accessed 20 March 2023].

Wittenberg, R., Knapp, M., Karagiannidou, M., Dickson, J. and Schott, J.M., 2019. Economic impacts of introducing diagnostics for mild cognitive impairment Alzheimer's disease

patients. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5(1), pp.382–387.

Xu, R., Yu, Y., Zhang, C., Ali, M.K., Ho, J.C. and Yang, C., 2022. Counterfactual and Factual Reasoning over Hypergraphs for Interpretable Clinical Predictions on EHR. *Proceedings of the 2nd Machine Learning for Health Symposium*, 28 November 2022, New Orleans and Virtual. [Online]: PMLR, pp.259–278.

Available from: https://proceedings.mlr.press/v193/xu22a.html [Accessed 21 May 2023].

Yuan, S., Wu, W., Ma, W., Huang, X., Huang, T., Peng, Mi., Xu, A. and Lyu, J., 2022. Body mass index, genetic susceptibility, and Alzheimer's disease: a longitudinal study based on 475,813 participants from the UK Biobank. *Journal of Translational Medicine*, 20(1), pp.417-428

Zhuang, D., Zhang, X., Song, S.L. and Hooker, S., 2021. *Randomness In Neural Network Training: Characterizing The Impact of Tooling* [Online]. arXiv. Available from: http://arxiv.org/abs/2106.11872 [Accessed 1 December 2023].

Zuo, Q., Lei, B., Shen, Y., Liu, Y., Feng, Z. and Wang, S., 2021. Multimodal Representations Learning and Adversarial Hypergraph Fusion for Early Alzheimer's Disease Prediction, In: Ma, H., et al. *Pattern Recognition and Computer Vision. PRCV 2021. Lecture Notes in Computer Science*, 29 October - 1 November 2021, Beijing. Cham: Springer, pp. 479-490.

Appendix A

CDR Score Definitions

	None 0	Questionable 0.5	Mild 1	Moderate 2	Severe
Memory	No memory loss or slight; inconsistent forgetfulness	Consistent slight forgetfulness; partial recollection of events; "benign" forgetfulness	Moderate memory loss: more marked for recent events; defect interferes with everyday activity	Severe memory loss, only highly learned material retained: new material rapidly lost	Severe memory loss, only fragments remain
Orientation	Fully oriented	Fully oriented but with slight difficulty with time relationships	Moderate difficulty with time relationships; oriented for place at examination; may have geographic disorientation elsewhere	Severe difficulty with time relationships; usually disoriented to time, often to place	Oriented to person only
Judgment and	Solves everyday problems and handles business and	Slight impairment in solving problems,	Moderate difficulty in handling problems,	Severely impaired in handling problems, similarities and	Unable to make judgments or solve problems

problem solving	financial affairs well; judgment good in relation to past performance	similarities and differences	similarities and differences; social judgment usually maintained	differences; social judgment usually impaired	
Community affairs	Independent function as usual in job, shopping, volunteer, and social groups	Slight impairment in these activities	Unable to function independently at these activities though may still be engaged in some; appears normal to casual inspection	No pretense of independent function outside the home; appears well enough to be taken to functions outside the family home	Appears too ill to be taken to functions outside the family home
Home and hobbies	Life at home, hobbies and intellectual interests well maintained	Life at home, hobbies and intellectual interests slightly impaired	Mild but definite impairment of functions at home; more difficult chores, and complicated hobbies and interests abandoned	Only simple chores preserved; very restricted interests, poorly maintained	No significant function in the home
Personal care	Fully capable of self-care	Fully capable of self-care	Needs prompting	Requires assistance in dressing, hygiene and keeping of personal effects	Requires much help with personal care; frequent incontinence

Morris (1993)

Appendix B

Supplemental Demographic EDA Diagrams

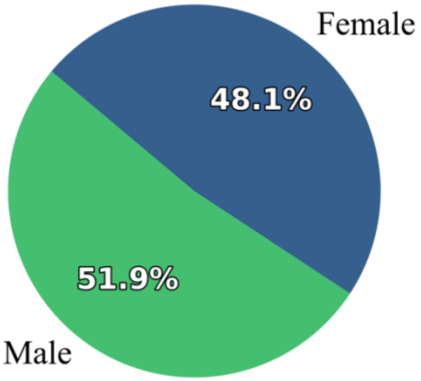

Figure 15: *Gender Distribution in the 3-year Cohort.* There is a slight bias towards male participants which may impact model learning.

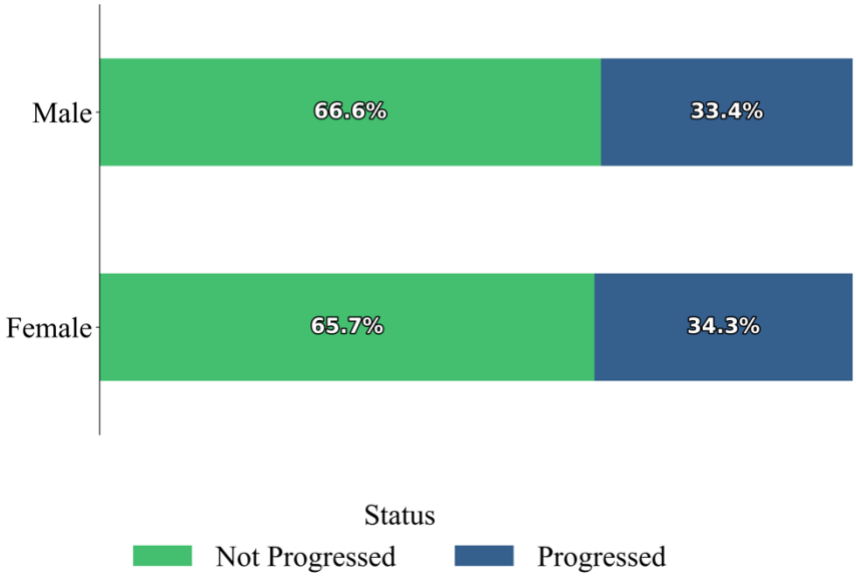

Figure 16: *Progression by Gender in the 3-year Cohort.* Shows the proportion of progression to dementia (global CDR score ≥ 1 within period) in men and women within the 3-years cohort. A slightly higher percentage of women progress.

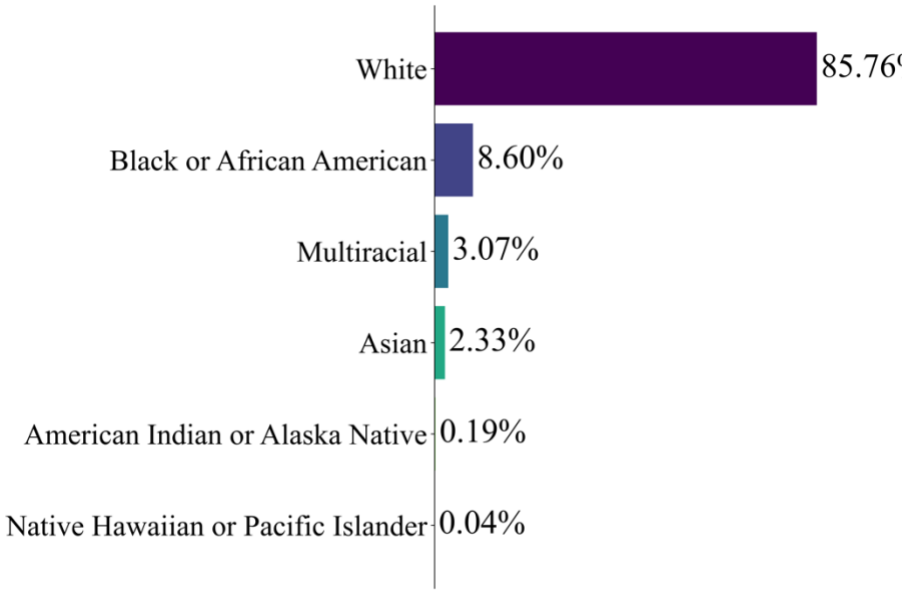

Figure 17*: Race Distribution in the 3-year Cohort.* The dataset is heavily biased to White participants with almost no representation for some races.

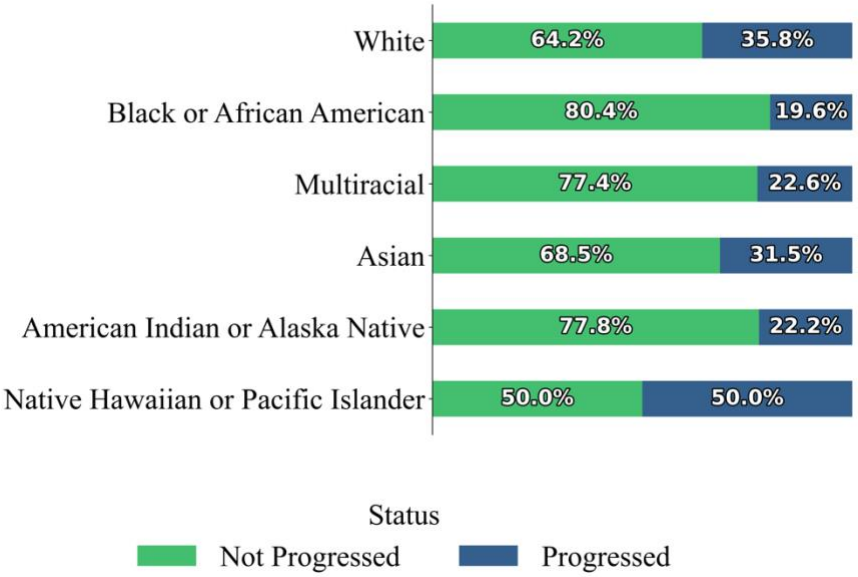

Figure 18: *Progression by Race in the 3-year Cohort.* Notably the progression rate of White and Asian participants is 13.2 and 8.9 percentage points higher than the next highest group (Multiracial), excluding Native Hawaiian or Pacific Islanders due to low representation.

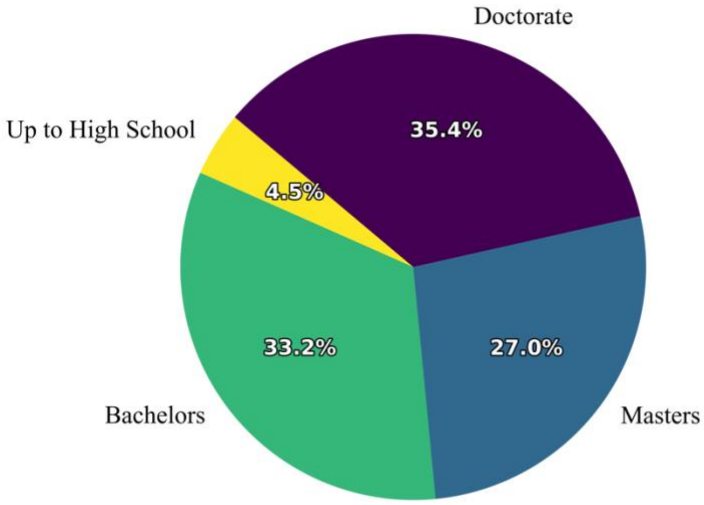

Figure 19: *Distribution of Education in the 3-year Cohort.* Notably, there is a very low percentage of participant with only a high school education and a very high percentage with a doctorate.

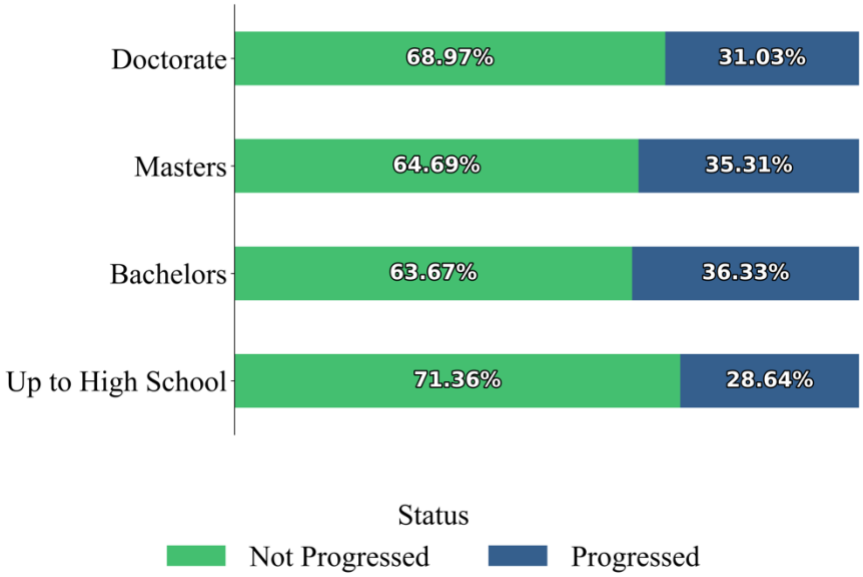

Figure 20: *Progression by Education Level in the 3-year Cohort.* We see a slight increase in progression rates as education level declines except for the up to high school group. The low progression rate in this group may be due to poor representation.

Appendix C

Supplemental Risk Factor EDA Diagrams

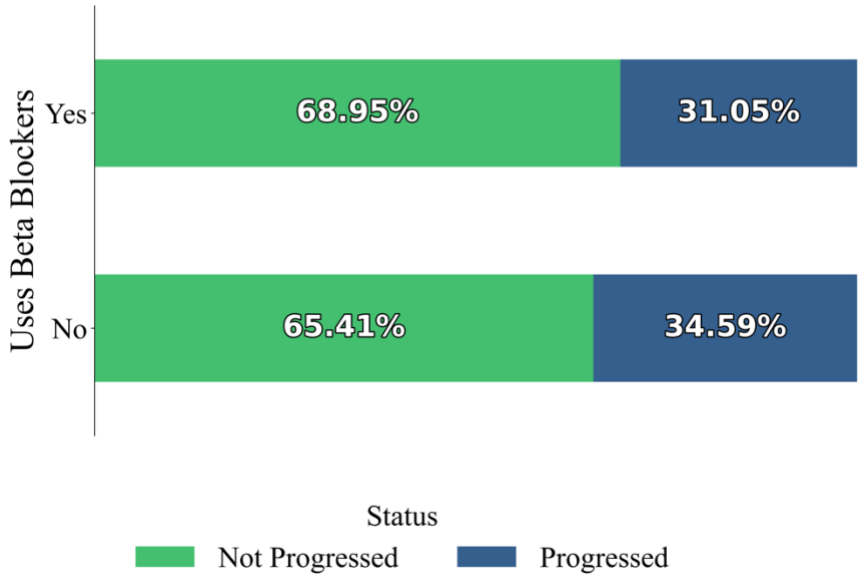

Figure 21: *Progression by Beta Blocker Use in the 3-year Cohort***.** We note a significantly higher rate of progression in the group which does not use beta blockers.

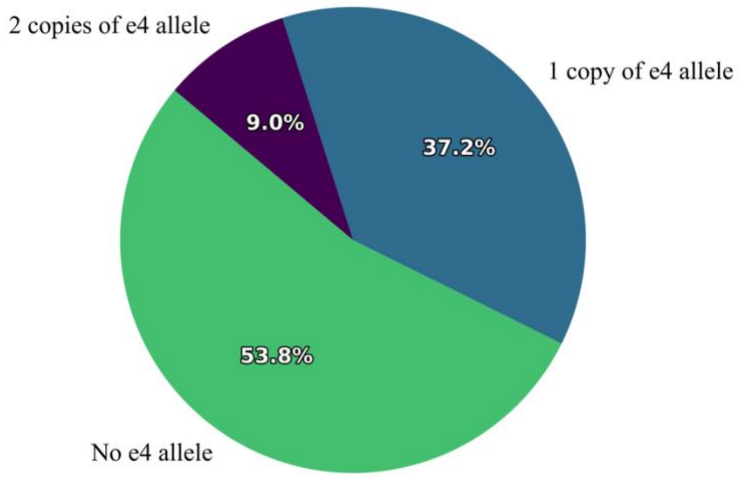

Figure 22: *Distribution of e4 Allele Count in the 3-year Cohort.* e4 alleles are a known risk factor in AD due to their role in increasing amyloid-beta build up. The group with 2 copies is the least represented by a large margin.

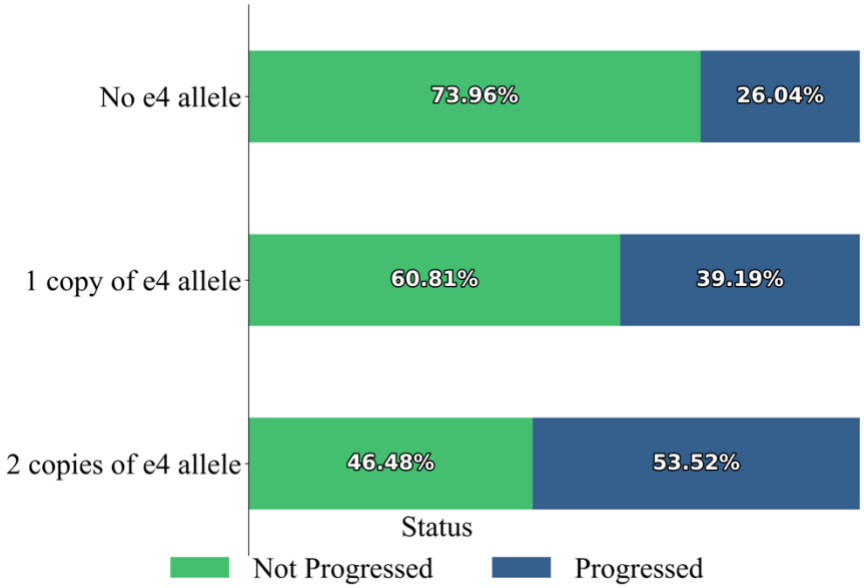

Figure 23: *Progression by e4 Allele Count in the 3-year Cohort.* Progression rates are over 50% higher for those with one copy and over 100% higher for those with two copies compared to those with none.

Appendix D

NACC Data Structure Overview

Columns	Section Descriptor
1 – 12	Administrative data
13 – 26	Milestone dates
27 – 36	Derived subject information
37 – 38	Telephone survey details
39 – 63	Subject demographics
64 – 85	Co-participant demographics
86 – 102	Subject family history
103 – 125	Subject medications
126 – 200	Subject health history
201 – 217	Physical health measurements
218 – 234	Hachinski Ishemic Score and cardio-vascular history
235 – 289	Unified Parkinson's Disease Rating Scale scores
290 – 299	CDR Scores
300 – 325	Neuropsychiatric Inventory Questionnaire
326 – 342	Geriatric Depression Scale Scores
343 – 352	Functional Activities Questionnaire
353 – 399	Physical / Neurological exam findings
400 – 459	Clinician judgement of symptoms
459 – 591	Neuropsychological Battery Scores, including MMSE
592 – 733	Clinician Diagnoses
734 – 766	Clinician-assessed Medical Conditions
767 – 792*	Genetic Summary Data (* available in a separate data sheet)

Table 7: *NACC Uniform Data Set Section Breakdown.* The NACC dataset is comprised of many columns of data covering a range of demographic and medical features.

Appendix E

Supplemental Experiment Results

WD	SAHs	DO	HCLs	VF_1	TF_1	Epoch
0	4	0	64	0.714	0.718	82
0	4	0	128	0.717	0.724	77
0	4	0	256	0.721	0.714	88
0	4	0.1	64	0.717	0.724	108
0	4	0.1	128	0.720	0.734	88
0	4	0.1	256	0.728	0.727	93
0	8	0	64	0.714	0.724	79
0	8	0	128	0.724	0.729	121
0	8	0	256	0.722	0.719	110
0	8	0.1	64	0.716	0.719	118
0	8	0.1	128	0.724	0.733	80
0	8	0.1	256	0.727	0.742	92
0.1	4	0	64	0.670	0.692	198
0.1	4	0	128	0.695	0.722	169
0.1	4	0	256	0.698	0.710	108
0.1	4	0.1	64	0.697	0.728	177
0.1	4	0.1	128	0.706	0.709	111
0.1	4	0.1	256	0.704	0.709	193
0.1	8	0	64	0.697	0.728	197
0.1	8	0	128	0.689	0.708	185
0.1	8	0	256	0.696	0.710	190
0.1	8	0.1	64	0.693	0.714	172
0.1	8	0.1	128	0.696	0.721	195
0.1	8	0.1	256	0.701	0.710	107

Table 8: *Grid search on EHNN Hyperparameters with Learning Rate of 0.001.* WD = Weight Decay; DO = Hypernetwork Dropout; SAHs = Self-attention Heads; HLCs = Hidden Layer Channels; VF_1= Validation Set F_1 Score; TF_1 = Test Set F_1 Score; Epoch = Epoch where the best validation F_1 score was achieved in that run. The best result was achieved with 0 WD, 4 SAHs, 0.1 DO and 256 HLCs. The test F_1 suggests no overfitting to validation set. We see overall worse performance and later convergence with WD, no benefit of extra SAHs, slight improvements with DO and best performance with high HCL.

WD	SAHs	DO	HCLs	VF_1	TF_1	Epoch
0	4	0	64	0.715	0.731	250
0	4	0	128	0.721	0.735	262
0	4	0	256	0.722	0.750	223
0	4	0.1	64	0.709	0.723	350
0	4	0.1	128	0.716	0.736	391
0	4	0.1	256	0.728	0.740	314
0	8	0	64	0.718	0.736	383
0	8	0	128	0.719	0.734	321
0	8	0	256	0.729	0.740	200
0	8	0.1	64	0.709	0.713	448
0	8	0.1	128	0.717	0.732	306
0	8	0.1	256	0.721	0.735	281
0.1	4	0	64	0.647	0.666	499
0.1	4	0	128	0.652	0.670	499
0.1	4	0	256	0.663	0.693	488
0.1	4	0.1	64	0.644	0.664	498
0.1	4	0.1	128	0.647	0.667	499
0.1	4	0.1	256	0.649	0.668	493
0.1	8	0	64	0.647	0.666	499
0.1	8	0	128	0.652	0.670	499
0.1	8	0	256	0.661	0.692	485
0.1	8	0.1	64	0.644	0.664	498
0.1	8	0.1	128	0.647	0.667	499
0.1	8	0.1	256	0.649	0.668	493

Table 9: *Grid search on EHNN Hyperparameters with Learning Rate of 0.0001.* WD = Weight Decay; DO = Hypernetwork Dropout; SAHs = Self-attention Heads; HLCs = Hidden Layer Channels; VF_1= Validation Set F_1 Score; TF_1 = Test Set F_1 Score; Epoch = Epoch where the best validation F_1 score was achieved in that run.

We see overall worse results compared to a learning rate of 0.01, with much later convergence overall. Again, we observe worse performance with WD and no significant difference in additional SAHs, in this case no clear difference in using DO and again improvement in increasing HLCs. Later convergence is likely due to training instability due to the higher learning rate.

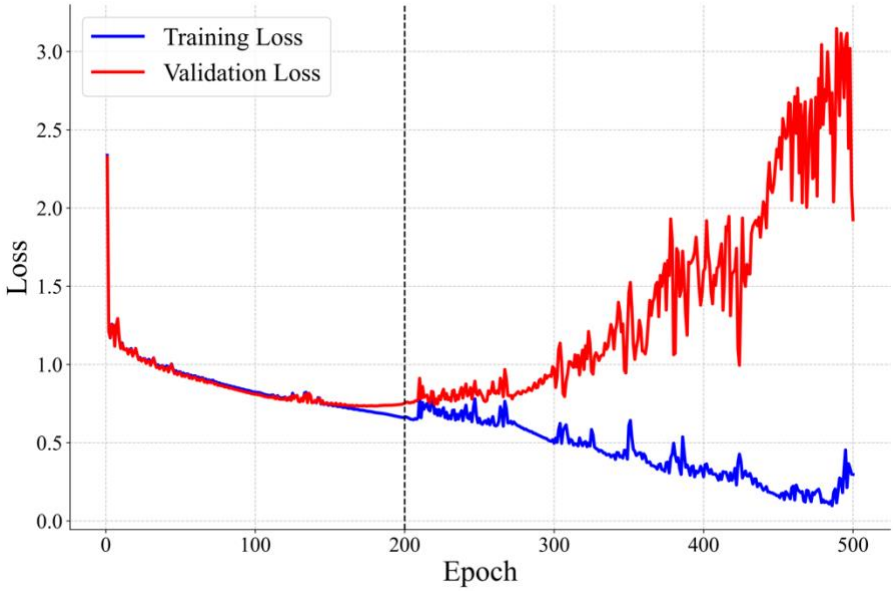

Figure 24: *Training and Validation Loss Curves for Best LR 0.0001 Grid Search Result.* Shows training and validation curves with LR 0.0001, 0 WD, 8 SAHs, 0 DO and 256 HLCs. The model shows overfitting at around 150 epochs and achieves the best result when significantly overfit.

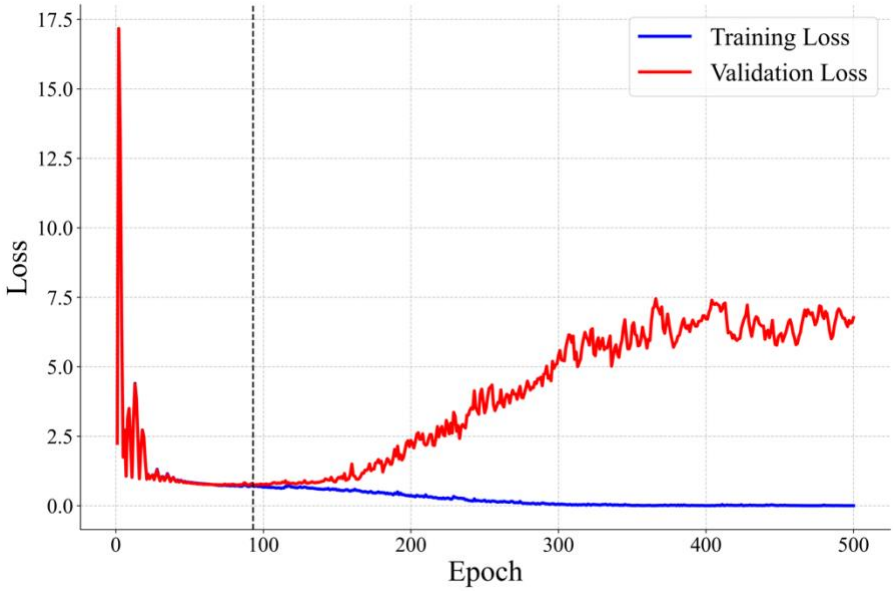

Figure 25: *Training and Validation Loss Curves for Best LR 0.001 Grid Search Result.* Shows training and validation loss curves with LR 0.001, 4 SAHs and 256 HLCs. Best score is achieved on epoch 93 denoted by the dotted line, after which overfitting is observed.

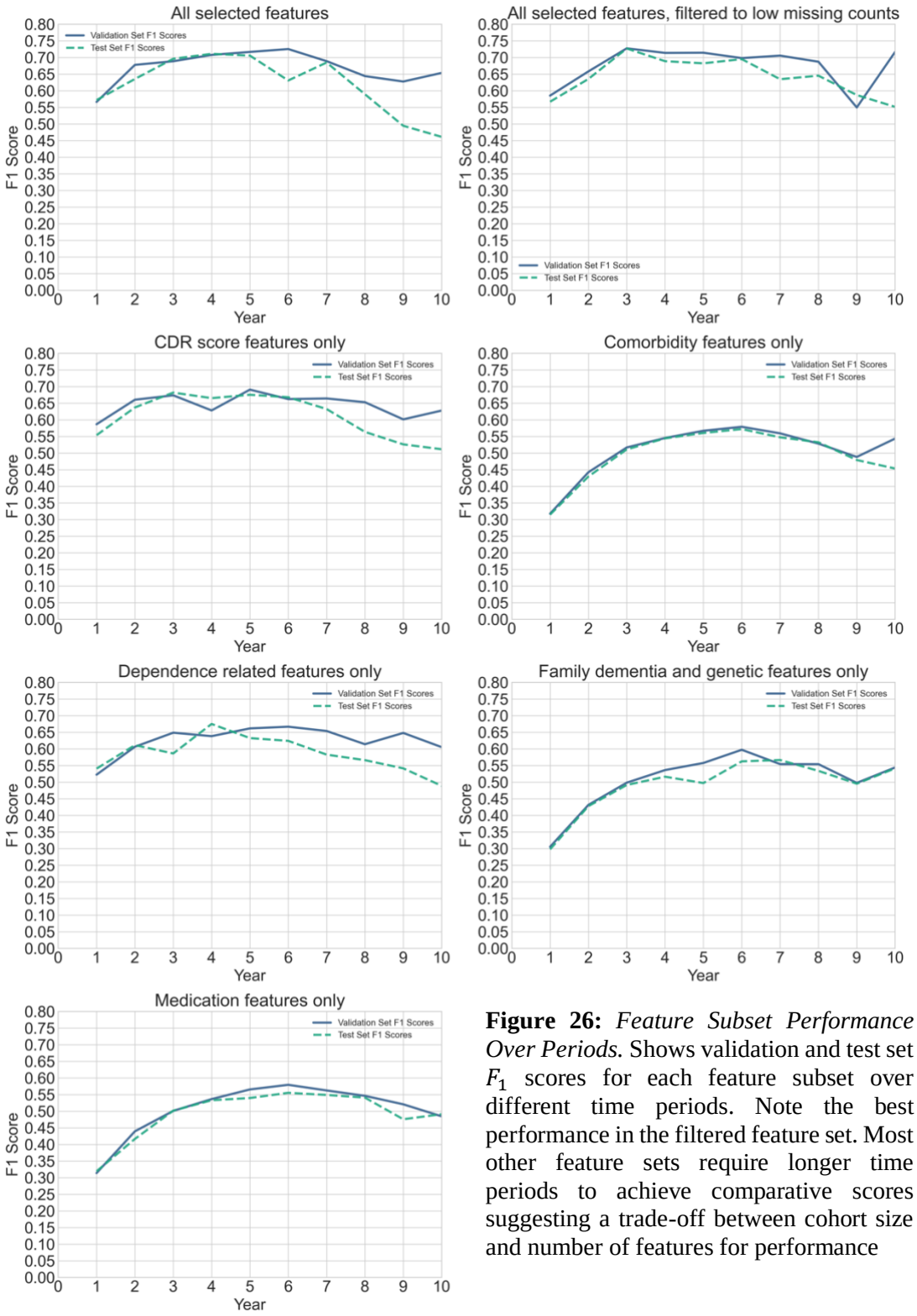

Figure 26: *Feature Subset Performance Over Periods.* Shows validation and test set F_1 scores for each feature subset over different time periods. Note the best performance in the filtered feature set. Most other feature sets require longer time periods to achieve comparative scores suggesting a trade-off between cohort size and number of features for performance

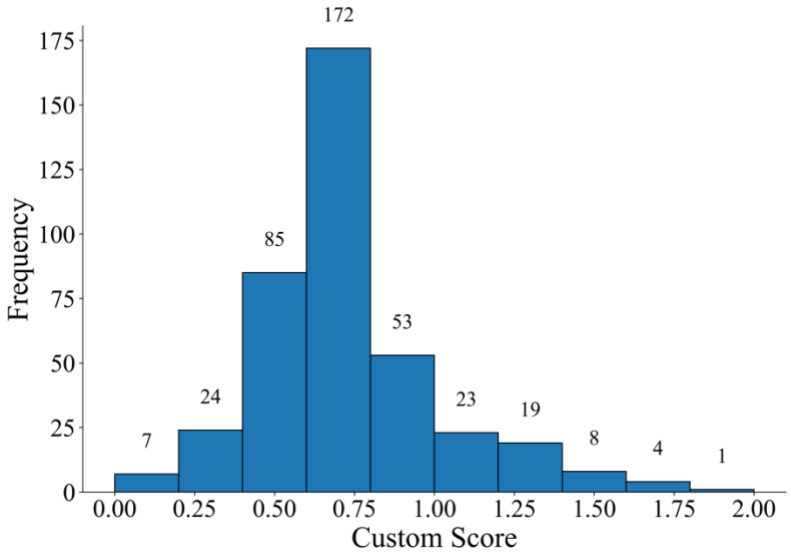

Figure 27: *Histogram of Hyperedge Score Distribution.* Of 396 hyperedges, 55 scored above 1, showing that a small number of hyperedges had the largest impact.