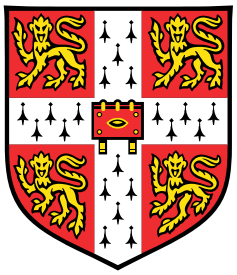
Graph Representation Learning for Child Mental Health Prediction



Ryan Crowley

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy in Machine Learning and Machine Intelligence

Churchill College August 2022

I dedicate this thesis to my parents, brothers, and partner, who have all supported me endlessly throughout the course of the degree.

Declaration

I, Ryan Crowley of Churchill College, hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains **14,879** words including appendices, footnotes, tables, and equations and has fewer than 150 figures.

Existing software: The software created in this thesis utilizes fundamental Python packages (PyTorch, NumPy, Scikit-Learn, Pandas, and Seaborn) to clean data, link datasets, and build foundational machine learning models. PyTorch-Geometric is utilized for creating and evaluating Graph Neural Network (GNN) models. For some of the GNNs, inspiration for implementation of code is derived from code in Rocheteau et al. [48]. All software code is implemented by me, unless otherwise specified.

Ryan Crowley
August 2022

Acknowledgements

I would like to first thank my supervisors, Professor Pietro Liò and Dr. Anna Moore, for their continual support and guidance throughout the course of my project. Both supervisors have provided invaluable knowledge that has contributed substantially to the work that I was able to produce. I am incredibly appreciative of the dual guidance regarding ML methodology and clinical knowledge from my supervisors and am grateful for the opportunity to have melded the two disciplines throughout the course of this thesis. I would also like to thank Emma Rocheteau, Katherine Parkin, Dr. Alisa Anokhina, and Dr. Efthalia Massou for their insightful commentary and work on assisting me throughout this project. Finally, I would like to thank my friends, family, and coursemates for helping me grow both academically and interpersonally.

Abstract

With rates of childhood depression continuing to rise, efforts to improve early identification of mental health are gaining increasing traction within psychiatry. Machine Learning (ML) methods for early identification of mental health provide the potential to scale these efforts broadly and equitably. However, for these ML methods to gain the trust of clinicians and key decision-makers within healthcare, they must demonstrate efficacy and interpretability.

This thesis thus seeks to characterize a new database for child mental health prediction and develop novel ML methodologies that provide both high-quality performance and interpretability. Specifically, this work focuses on creating patient graphs using novel similarity metrics and applying Graph Neural Networks (GNNs) to these patient graphs for child mental health prediction. Results indicate that this approach is superior to standard Neural Network methods that do not take advantage of graphical structures. In particular, measuring patient similarity by calculating distance between latent space embeddings of patient data derived from an autoencoder network shows impressive performance in clustering individuals with mental health diagnoses.

To comprehensively evaluate the performance of these methods, I conduct additional analyses including assessments of algorithmic fairness, clinical interpretability of Graph Attention Network models, and methods for handling missing data. Overall, this thesis seeks to unify work within psychiatry and GNN development with findings that are broadly applicable for both fields.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

ACEs Adverse Childhood Experiences

AI Artificial Intelligence

AUPRC Area Under the Precision Recall Curve

AUROC Area Under the Receiver Operating Characteristic Curve

FP Feature Propagation

GAT Graph Attention Network

GNN Graph Neural Network

ML Machine Learning

MLP Multilayer Perceptron

MPNN Message Passing Neural Network

NN Neural Network

NPV Negative Predictive Value

PPV Positive Predictive Value

RNN Recurrent Neural Network

SAGE GraphSAGE

SAIL Secure Anonymised Information Linnkage

TNR True Negative Rate

TPR True Positive Rate

Chapter 1

Introduction

1.1 Thesis Structure

The narrative arc of the thesis will follow the structure described below.

Chapter 1 will discuss the overarching research goals, provide necessary context for understanding the work as a whole, and illuminate the primary contributions that will be presented.

Chapter 2 will then discuss prior research on child mental health as well as the technical background of Graph Neural Networks (GNNs) and other associated methods that will be investigated in the foregoing analyses.

Chapter 3 will explore the Secure Anonymised Information Linkage (SAIL) Database, focusing on characterizing the specific conglomeration of datasets analyzed in this work.

Chapter 4 will then discuss the development and analysis of a multitude of different similarity metrics that can be used for homophilic graph construction. These similarity metrics will be applied for construction of patient graphs.

Chapter 5 will explore different GNN formulations that use patient graphs for child mental health prediction. Performance will be compared to baseline NN approaches and analyses of model fairness will be conducted.

Chapter 6 will expand upon the GNN methods analyzed in Chapter 5 by applying Feature Propagation (FP) for missing data and analyzing the clinical interpretability potential of Graph Attention Networks (GATs).

Chapter 7 will unify the preceding chapters, discuss the limitations of the methods and modelling approaches proposed, and provide recommendations for future work.

1.2 Overall Research Goals

This thesis contributes to the overall research conducted and coordinated by Dr. Anna Moore, a clinical psychiatrist focused predominantly on developing early identification tools for child mental health by linking data from diverse sources. One of the overarching goals of her research is to create the Cam-Child database, a linked administrative database that will contain comprehensive health, educational, and social care data for all children within Cambridgeshire and Peterborough. The Cam-Child database is intended to be used across a range of diseases and for both exploratory research and interventions. This dataset will be a valuable step towards improved monitoring and treatment of adolescent mental health in the region and could inspire similar projects in other locations.

To obtain a head-start on building analysis pipelines and assessing the performance of various models for predicting child mental health while the Cam-Child dataset is being constructed, access to the Secure Anonymised Information Linkage (SAIL) Databank [23] was secured. Similar to the Cam-Child dataset, the SAIL Databank comprehensively covers all individuals within a given area and contains health, social care, and educational data. However, the datasets differ in that the SAIL Databank is specific to Wales, is used only for research purposes, and does not allow for the deanonymization of data to stage interventions. Consequently, by developing new methodologies and model frameworks to predict child mental health within the SAIL Databank, the goal is to prepare these methodologies for potential future use in real-world applications through analysis of the Cam-Child dataset. Given the ultimate goals of the project, the primary clinical questions that inspire the methodological advancements proposed in this research are:

1. Can we create Machine Learning (ML) models using diverse data from linked datasets to accurately identify children with a mental health diagnosis?

2. How can we improve the interpretability of these models to be more useful in a clinical or administrative setting for staging mental health interventions?

To ensure that the model and findings are clinically useful, these clinical questions will directly inform the methodologies developed. Throughout the course of this work, I will seek to develop and validate novel ML methodologies that incorporate educational, social care, and healthcare data to answer these two key clinical questions. This work focuses on

GNNs specifically as an approach to incorporate additional useful clinical information into the model for improved prediction.

1.3 Main Contributions

This work seeks to make advancements relevant to both the psychiatry literature and the GNN literature. The primary contributions of the work are as follows:

1. Characterization of a novel dataset integrating information from diverse sources for use in child mental health prediction.
2. Development of novel similarity metrics for homophilic graph construction including use of an autoencoder network.
3. Evaluation of performance and fairness of GNN models in the novel domain of child mental health prediction.
4. Application and evaluation of Feature Propagation for addressing missing data problems and incorporating information regarding the outcome labels of neighbors into model prediction.
5. Assessment of the utility of Graph Attention Network attentional weights for adding interpretability for clinical decision-making.

A graphical abstract showing the main structure of the thesis is shown in Figure 1.1 demonstrating the focus on creation of similarity metrics, development of GNN models, and evaluation of the fairness and interpretability of GNN models.

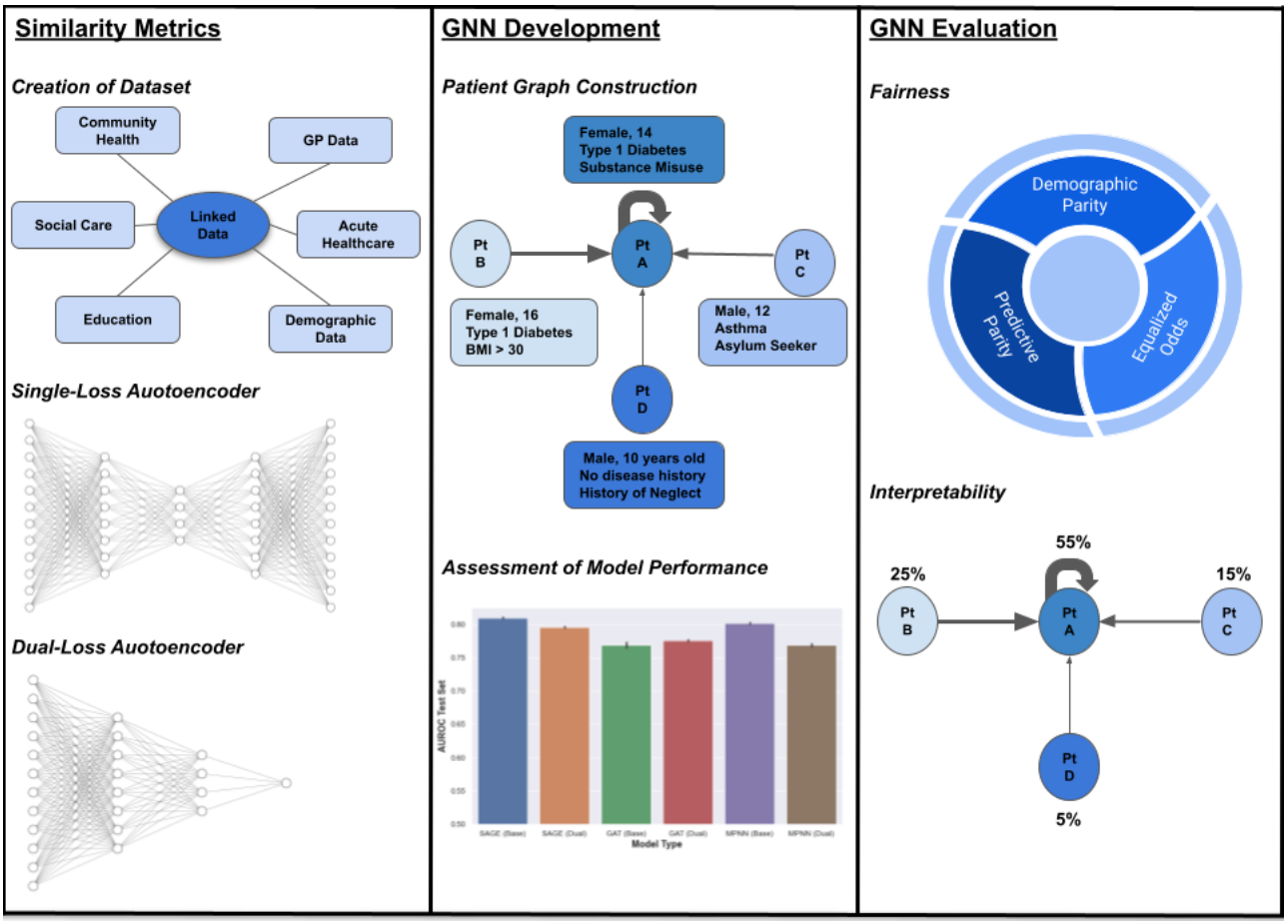


Fig. 1.1 Overview of the Main Approaches Adopted in this Thesis

Having discussed the overarching research goals and the main contributions of the work, I will focus in Chapter 2 on the technical background that underpins the work conducted in this thesis.

Chapter 2

Background

2.1 Child Mental Health

Incidence and prevalence of childhood mental health disorders are on the rise in the UK. Measurements of the extent of the problem vary, but a recent report places the prevalence of childhood mental health disorders at approximately 10% [34]. This increase in child mental health disorders likely stems from a confluence of factors including widening income inequality, changes in family environments, social media usage, and increased pressure within school settings [3]. The WHO states that one out of four people will be affected by mental health issues in their lives and that depressive disorders will become the second leading cause of global disease burden trailing only ischemic heart disease [53]. The rise of child mental health disorders is a troubling phenomenon that deserves time, attention, and resources.

Diagnosing children with mental health issues is particularly difficult because children may experience different symptoms depending on their age, children may be less able to explain their feelings and behaviors, and typical childhood development is inherently a process that involves change, making it difficult to separate development from disorder [46]. Despite the difficulties of childhood diagnosis of mental health issues, early detection of mental health disorders is crucial as it allows for effective interventions before problems become more severe and debilitating [45]. Furthermore, psychopathology that children experience early in life can lead to negative outcomes that affect them throughout adolescence and into adulthood. Early identification of child mental health disorders should be an integral part of our systemic response to the impending mental health crisis.

The literature on child mental health focuses on Adverse Childhood Experiences (ACEs), which encompass both direct harms such as physical abuse and neglect and indirect harms

such as substance misuse and parental conflict [20]. Cumulative experience of ACEs is tied to behavioral problems in children [34] and a variety of negative health outcomes [20]. The effect of ACEs on the body is holistic; in conjunction with other factors like socioeconomic status, ACEs even have a long-lasting impact on biological regulation [41]. Since the impact of some of these ACEs and other documented risk factors can be mitigated by early interventions [55], early intervention programs are of incredible importance. ACEs and adverse mental health outcomes are inextricably linked to healthcare disparities, and there are substantially higher levels of mental health diagnoses in areas with higher social deprivation and lower average socioeconomic status [34].

2.1.1 Clinical Prediction Models

Despite the increasing burden of mental health on healthcare systems, the growth in the number of psychiatrists and healthcare professionals trained in mental health is significantly outpaced by the number of people suffering [53]. Hence, clinical prediction tools provide the potential to ameliorate the growing child mental health epidemic by aiding in the early identification of mental health disorders. More broadly, ML approaches are becoming increasingly commonplace in the mental health sphere. Some recent applications include precision psychiatry applications using pharmacogenomics [31], monitoring of user activity on social media [28], and the use of ML algorithms for improvement of virtual medical assistants who provide support to individuals with mental health problems [2, 59] (spurred on by research indicating that individuals may be more comfortable disclosing sensitive information to a computer system than with a person [35]). One of the primary focuses of this body of research is the application of ML to diagnosis tasks with some of the most commonly assessed diseases including schizophrenia, depression, bipolar disorder, and anxiety [9].

Nevertheless, despite the promise of predictive risk tools and extensive research within the area, no tools are yet available for clinical use within psychiatry for detection of child mental health disorders [51]. More broadly, as of the writing of this thesis, there are currently no FDA-cleared Artificial Intelligence (AI) applications for psychiatry [29]. Further, the overall quality of existing mental health prediction models is generally poor. In a Cochrane review of the literature, prognostic models for prediction of relapse and/or recurrence of major depressive disorder in adults were assessed. The overall results indicate that few high quality models exist and most models are at high risk of bias [42]. This discrepancy between the vast potential for ML applications and the corresponding lack of improvement in patient outcomes has been dubbed the "AI chasm." A predominant reason that the AI chasm exists is that poor evaluations of model performance are rampant in medical AI; evaluations are

typically conducted via internal validation, without proper safeguards, and using methods that may overestimate model performance [10]. Additionally, ML studies often focus on conducting analyses of accuracy with incredibly few studies assessing the more crucial metric: improved patient outcomes [13].

The reasons underlying the AI chasm within medicine are often magnified within psychiatry. Datasets used for AI applications to psychiatry are often small and validation is typically conducted internally without appropriate safeguards [9]. Additionally, psychiatry models suffer from low generalizability and interpretability, assessments are predominantly conducted in homogeneous populations in affluent countries, and external validations of model performance are exceedingly rare [40, 4, 29]. Finally, model performance is often poor due to a lack of knowledge regarding the underlying pathology of psychiatric diseases and a lack of understanding of the assumptions of ML models [11]. These issues relating to clinical and technical knowledge are likely compounded by faulty or missing communication between ML scientists and clinicians.

Most of the ML research within psychiatry has been conducted using standard, baseline models such as Random Forest and Support Vector Machines, which do not take advantage of the inherent graphical nature of health data [9]. Hence, this project seeks to address the numerous problems within the field of ML algorithms for psychiatry by developing and validating novel graphical ML approaches to create a diagnostic tool for child mental health prediction. As in many other AI applications, there exists the potential for AI to exacerbate issues related to fairness, transparency, privacy, security, and accountability. In this work, I seek to address many of the limitations prevalent in the literature by focusing on interpretability of GNNs, assessments of algorithmic fairness, and taking an integrated approach that includes perspectives from both ML scientists and clinicians. Fairness and transparency, specifically, will be discussed in detail in Chapters 5 and 6 in order to help counteract the tendency of AI algorithms to reinforce historical patterns of discrimination and systemic bias [30]; however, the topics of privacy, security, and accountability will be left for future work focused upon implementation.

2.2 Graph Neural Networks (GNNs)

Graph Neural Networks (GNNs), the methodological focus of this thesis, are a subclass of Neural Networks designed to perform inference on data presented in a graphical form. GNNs use feature propagation and aggregation methods to combine feature information and graph structure in order to learn better representations of data [33]. Different GNN models

diverge in their approaches to neighborhood aggregation and node transformation methods. In this thesis, I choose to concentrate on three standard GNNs: GraphSAGE (SAGE) [18], Message Passing Neural Networks (MPNN) [15], and Graph Attention Networks (GAT) [56]. Detailed theoretical treatment and derivations of these models can be found in their accompanying papers cited above. In relation to health data specifically, GNNs are gaining increasing traction as a method for dealing with heterogeneous Electronic Health Record (EHR) data [32, 44, 1]. Other common ML approaches to EHR data include the use of Long Short-Term Memory (LSTM) Networks or other similar Recurrent Neural Network (RNN) models that incorporate temporal information [52], though these models will not be the focus of this work due to the lack of granular time data in the SAIL Databank. I conjecture that adapting graphical approaches for use within the SAIL database could help to leverage the rich information inherent within the clinical data to achieve performance superior to that obtained by more traditional ML methods utilized for diagnosis in psychiatry.

Most research on graphs for EHR data has focused on using knowledge graphs to inject clinical knowledge into the model. For instance, GRaph-based Attention Model (GRAM) supplements EHR data with the hierarchical information present within medical ontologies [8]. Alternative approaches include the use of a variationally regularized encoder-decoder graph network that implicitly learns the graph structure via a self-attention mechanism [62] and Knowledge-Based Attention Model (KAME), which uses a medical ontology to learn embeddings that are later fed into an RNN for diagnosis prediction [36]. All of these methods are centered around the popular approach of knowledge graphs; in contrast, patient graphs are rarely explored within the literature. To the best of my knowledge, the only other examples of patient graphs are the use of a patient affinity graph for missing data imputation in prediction of in-hospital mortality [38] and the creation of a patient similarity graph using the hierarchical information present within ICD-10 codes for prediction of length of stay using a hybrid LSTM-GNN model [48].

The methods proposed in this thesis differ from existing work on GNNs in that this work focuses on the understudied task of child mental health prediction, incorporates diverse data extending beyond solely healthcare data, explores novel similarity metrics for patient graph construction, assesses data imputation methods, and conducts analyses of graph interpretability. In the ensuing chapter, I will first focus upon characterizing the SAIL dataset to provide a deeper understanding of the underlying data to be used for mental health prediction.

Chapter 3

Characterization of SAIL Database

3.1 Motivation

This chapter delves into description and characterization of the SAIL Database. This dataset, containing diverse data on children from Wales, will be used for development and evaluation of the methods applied in this thesis. I explore relevant aspects of the dataset including prevalence of mental health outcomes, amount of missing data, and demographic characteristics.

3.2 SAIL Database Overview

3.2.1 Baseline Cohort

The SAIL Databank contains a plethora of information about the population of Wales across a variety of databases. The Adolescent Data Platform (ADP) includes information relating to youth within Wales with databases that contain demographic data, acute care data, NHS outpatient data, General Practicioner (GP) data, educational attainment data, substance misuse data, and other additional data. The baseline cohort consists of >1,000,000 individuals aged 0-17 years within the years 2013 to 2020, plus any retrospective and subsequent data relating to these young people. To the best of my knowledge, no one has utilized the SAIL database to predict mental health disorders in children, so this research could provide incredibly important insights relevant to child psychiatry.

With the help of Dr. Yasmin Friedman and others, I linked 18 different datasets containing pertinent information. This process included utilizing demographic information and local identifiers from the 18 different datasets to link all individuals to a unique Anonymous Linking Field (ALF), an anonymous identifier assigned to each child in the cohort. Given the

diverse nature of local identifiers for the social care, healthcare, and education datasets, this was an intensive process. Once completed, the data from each of the different datasets could be collated for prediction. All 18 datasets were then grouped into the following categories for easier interpretation: demographics, education, social care, GP, community healthcare, and acute healthcare. The groupings of the different datasets, their names, and their abbreviations are shown in Figure 3.1.

Group Name	Individual Data Sources
Demographics	Welsh Demographic Service Dataset (WDSD) Annual District Birth Extract (ADBE) Annual District Death Extract (ADDE)
Education	Pre16 Education Attainment (EDUW)
Social Care	Child In Need Wales (CINW) Children Receiving Care and Support (CRCS) Looked After Children (LACW)
GP	GP Primary Care – Audit (WLGP)
Community Healthcare	National Community Child Health (NCCH) Maternity Indicators Dataset (MIDS) NHS Hospital Outpatients (OPDW) Outpatient Referrals from Primary Care (OPRD) Wales Results Reporting Service (WRRS) Substance Misuse Dataset (SMDS) NHS 111 Call data (NHSO)
Acute Healthcare	Patient Episode Database for Wales (PEDW) Emergency Department Dataset (EDDS) Critical Care Dataset (CCDS)

Fig. 3.1 Schematic of SAIL Dataset Groupings

Together, these datasets encapsulate a diverse array of informative variables related to mental health that are often analyzed independently with few efforts seeking to unify this information into a single, linked dataset. The different databases also vary in relation to the time periods that they cover. Figure 3.2 shows the temporal coverage of each of the databases.

Although some databases existed before 1996, our cohort only consists of individuals born in 1996 or later, so the lower-bound for all databases is 1996. Moreover, the most recent information from the datasets is from early 2022 so the upper bound is set accordingly.

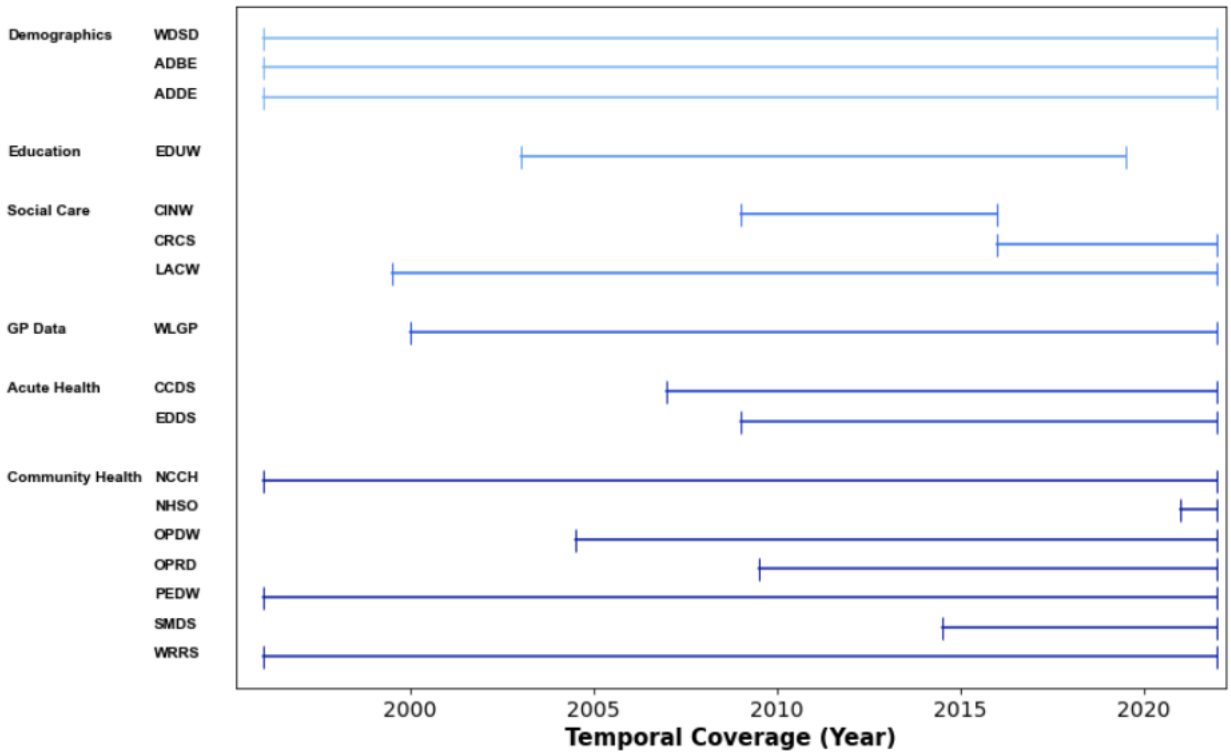


Fig. 3.2 Temporal Coverage Afforded by SAIL Datasets

Examining the information in Figure 3.2, we find that most datasets are still collecting data through the current day. The two notable exceptions are the education dataset (EDUW) and the Children in Need Wales Dataset (CINW). The EDUW dataset, which has not collected new data since 2019, was likely affected by school closings due to COVID-19, which hindered the ability to collect education data with a high level of precision. Moreover, the CINW database became the CRCS dataset following the enactment of the Social Services and Well-being (Wales) Act in April 2016. The CINW and CRCS datasets provide similar information relating to children in need with substantial overlap.

3.2.2 Delphi Study Identifying Risk Factors

In order to improve the understanding of the diverse risk factors relevant to mental health prediction, Katherine Parkin and others undertook a rapid review and Delphi process that identified over 250 variables grouped into seven domains: social and environmental, behavioural,

education and employment, biomarkers, physical health, psychological and mental health, and patterns of service use. A special eighth domain highlighting risk factors especially pertinent for underserved populations was also included. The risk factors from the Delphi study were then mapped by Katherine Parkin to the variables in the SAIL databank for use as clinical prediction features. This process included cross-referencing all risk factors from the Delphi with all variables from SAIL in order to identify which risk factors were directly measurable or derivable from the variables present within SAIL, leading to the creation of the list of variables used in these analyses.

3.2.3 Social Care Datasets

In order to keep analysis tractable, I have focused my ML development methods on individuals with social care data available. The social care datasets are particularly helpful because they contain a variety of pertinent risk factors (substance misuse, disability status, etc.) combined with lower levels of data missingness. As mentioned above, the Children In Need Wales (CINW) Dataset precedes the Children Receiving Care and Support (CRCS) Dataset, and both datasets contain information for children who have a care and support plan. Looked after children also have a care and support plan so they comprise a subset of this population.

There are 31,423 individuals with data present within CINW, 26,041 individuals with data present within CRCS, and 10,772 individuals in both datasets. In total, there are 46,704 unique individuals with social care data who will form the cohort explored in this work. In situations where individuals have data present in both datasets, all applicable data is used for prediction. Variables were also included from the other 18 datasets, as long as the variable of interest had missing values for less than 70% of the individuals within the social care cohort.

3.2.4 Time Series Feasibility and Dataset Splitting

Many EHR methods have shown success by utilizing LSTM or RNN model structures to model time-series data [17, 48, 50]. Hence, I critically assessed the possibility of utilizing methods that explicitly model time for child mental health prediction. The variables included within the model all contain time stamps which indicate when the individual experienced that risk factor. However, the social care data stems from an annual census. Thus, this lack of granular time series data led me to avoid explicitly modelling time within the model. In addition, the frequency of many timestamps is irregular, the time stamps may not correspond to when risk factors actually occurred, and there is some missingness for time stamps. Hence,

although I lose some important temporal information by not explicitly modelling time, this is the only viable decision and provides additional flexibility in our computational approach that will be explored in later sections.

To follow with standard conventions and ensure that model performance is not overestimated, the dataset is randomly split such that 70% of individuals fall into the training dataset, 15% into the validation dataset, and 15% into the testing dataset. This corresponds to 32,692 individuals in the training set, 7,006 individuals in the validation set, and 7,006 individuals in the test set.

3.3 Dataset Characterization

3.3.1 Demographic Information

After linkage and extensive data cleaning was completed, characterization of the dataset ensued to gain a deeper understanding of the information present within the dataset. Information from the Welsh Demographic Service Dataset (WDSD) was utilized to conduct analyses of the biological sex (Table 3.1), age (Table 3.2), and ethnicity (Table 3.3) of all individuals within the cohort. Age is calculated for the entire cohort as of January 1st, 2022.

Table 3.1 Distribution of Biological Sex within Cohort

Biological Sex	**# of Individuals**	**Percent of Cohort**
Male	25,179	53.91
Female	21,525	46.09
Total	**46,704**	**100**

Table 3.2 Distribution of Age within Cohort

Age	# of Individuals	Percent of Cohort
4-7	1644	6.21
8-11	7,926	16.97
12-15	11,124	23.82
16-19	11,455	24.53
20-23	9,355	20.03
24-27	3,989	8.54
Total	**46,704**	**100**

Table 3.3 Distribution of Ethnicity within Cohort

Ethnicity	# of Individuals	Percent of Cohort
Asian	855	1.83
Black	542	1.16
Mixed	1,317	2.82
Other	3,122	6.68
White	40,868	87.50
Total	**46,704**	**100**

Table 3.1 shows that the cohort consists of 46,704 individuals with a larger percentage of males than females with social care data. I will explore whether this discrepancy in data quantity for different genders affects model performance in Chapter 5. In Table 3.2, we can see that the age bands containing the highest number of individuals are 12-15 years and 16-19 years. The youngest individual within the social care dataset is 4.69 year old, the oldest individual within the dataset is 26.02 years old, the median age of individuals within the dataset is 16.49 years, and the mean age of individuals within the dataset is 16.41 years. It is important to note that many individuals within the cohort are no longer considered youth as the CINW dataset has been in existence since 2010. However, although the individual may no longer be a minor, their associated social care data and other pertinent information utilized in the model are taken from their childhood years. Table 3.3 demonstrates that the dataset is comprised predominantly of White children with non-White children comprising just 12.50% of the dataset. This lack of ethnic diversity is important to consider when assessing model

generalizability, and the impact of ethnicity on model performance and health equity will be explored in Chapter 5.

3.3.2 Missing Data

In order to gain greater insight into methodological approaches suitable for this cohort, an initial assessment of missing data was conducted. Missing data was calculated both on an individual level and on a variable level. Nonsense values such as a birth weight of 0.0 grams or a maternal age of 215 years were classified as missing data. The missing data on a per individual level is shown in Figure 3.3, where the y-axis is the count of individuals within that specific bin.

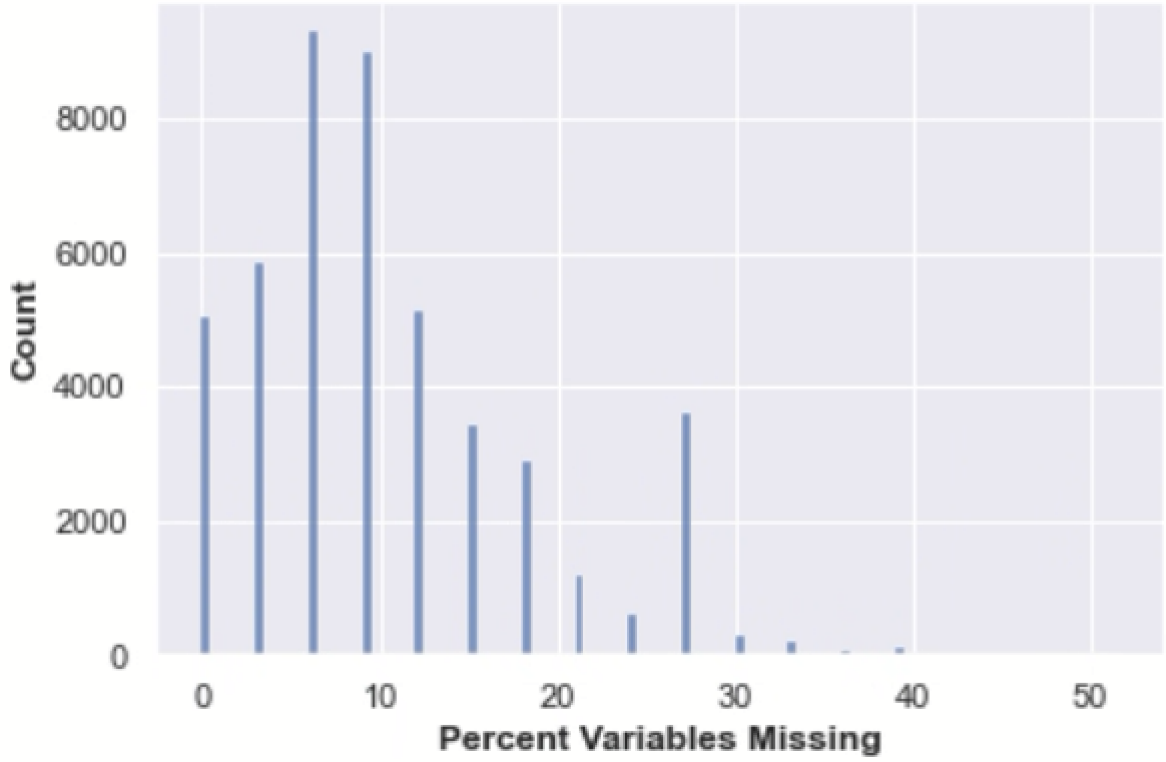


Fig. 3.3 Percentage of SAIL Variables with Missing Data Per Person

Figure 3.3 shows that individuals typically have a relatively small amount of missing data, and most individuals have missing data for less than 20% of variables included within the model. The mean percent of variables missing for an individual is 10.35%, the median is 9.09%, the minimum is 0.00%, and the maximum is 51.51%. This intermediate level of missingness indicates that the dataset likely has data that is of sufficient quality for

meaningful ML prediction. Missing data is also assessed on the level of variables in Figure 3.4, where the y-axis indicates the number of variables within that specific bin.

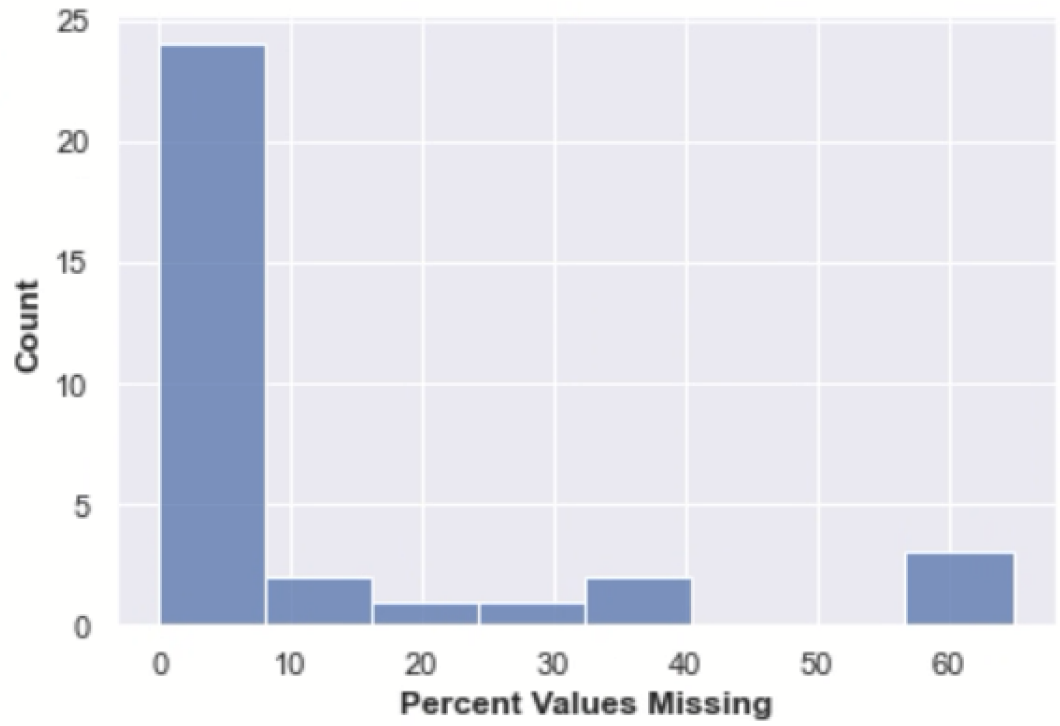


Fig. 3.4 Percentage of Missing Values Per SAIL Variable

Examining the percentage of missing values per variable in Figure 3.4, we can identify that many variables have less than 10% of their values missing. These variables with low missingness predominantly correspond to variables from the social care datasets and demographic variables like age and gender. Other variables have substantially larger missingness including some variables that are missing data for over 50% of individuals within the cohort. These variables with large amounts of missingness predominantly relate to health variables like birth weight and Apgar scores. The mean percent of values missing per variable is 10.36%, the median is .95%, the minimum is 0.00%, and the maximum is 64.82%. Techniques for addressing this missing data are explored in Chapter 6.

3.3.3 Description of Variables

Variables from the social care CINW/CRCS datasets as well as variables from the other datasets were cleaned and prepared for use. Categorical variables are listed in Table 3.4 and continuous variables are listed in Table 3.5.

Table 3.4 SAIL Categorical Variables Included in Modelling Approach

Variable Name	Dataset
Asylum Seeker Status	CINW/CRCS
Autistic Spectrum Disorder Status	CINW/CRCS
Breastfeeding Status (Birth)	MIDS
Breastfeeding Status (8 Weeks)	MIDS
Category of Need	CINW/CRCS
Child Protection Register Status	CINW/CRCS
Dental Check Status	CINW/CRCS
Disability (Memory)	CINW/CRCS
Disability (Mobility)	CINW/CRCS
Disability (None)	CINW/CRCS
Disability (Sensory)	CINW/CRCS
Free School Meal Status	EDUW
Gender	WDSD
Health Surveillance Checks Status	CINW/CRCS
Immunizations Status	CINW/CRCS
Labour Onset	NCCH
Looked After Child Status	CINW/CRCS
Maternal Smoking	NCCH
Parenting Capacity (Domestic Abuse)	CINW/CRCS
Parenting Capacity (Learning Disabilities)	CINW/CRCS
Parenting Capacity (Mental Health)	CINW/CRCS
Parenting Capacity (Physical Health)	CINW/CRCS
Parenting Capacity (Substance Misuse)	CINW/CRCS
School Exclusion Category	EDUW
Substance Misuse	CINW/CRCS
Urban/Rural Status	ADBE
Youth Offending Status	CINW/CRCS

Table 3.5 SAIL Continuous Variables Included in Modelling Approach

Variable Name	Dataset
Age	WDSD
Apgar 1-Minute	NCCH
Apgar 5-Minute	NCCH
Birth Weight	NCCH
Gestation Age	NCCH
Welsh Index of Multiple Deprivation	WDSD

Since I focused solely on individuals with social care data, many of the variables come from the social care datasets because these variables are more likely to meet the minimum criteria for amount of complete data necessary for inclusion. Some of the variables used for modelling relate to an individual's early life and child development (maternal smoking status, birth weight, etc.), while some relate to basic demographic data (sex, age, urbanicity, etc.), and others relate to difficult childhood experiences (disabilities, domestic abuse, substance misuse, etc.). Taken together, these variables contain diverse information relating to risk factors relevant at many different stages of an individual's life course.

3.3.4 Diagnosis and Operation Data

Some of the richest clinical information contained in the SAIL database is found within diagnosis codes and operation codes. Diagnosis codes within SAIL are predominantly found in the PEDW and OPDW datasets and follow the format of ICD-10, a medical classification list that contains hierarchical codes for diseases, signs and symptoms, and social circumstances. Operation codes are also found within SAIL in the PEDW and OPDW datasets and follow the format of the OPCS-4 Classification of Interventions and Procedures, a medical classification list that contains hierarchical codes for operations. Any codes that begin with "F" within ICD-10 are removed because these codes relate to mental health and/or neurological disorders. Read codes from GP data are also incredibly information-rich, but their heterogeneity and variability make them quite difficult to use, so they are not explored in this study.

Despite the importance of this clinical information, it is difficult to incorporate diagnosis and operations data into clinical models due to the high dimensionality and sparsity of this data; consequently, rare diseases and operations often do not have sufficient data for accurate

prediction. Within this cohort, there are 6,325 unique diagnosis codes and 4,080 unique operation codes. 79.43% of individuals have a diagnosis listed within their clinical data and 48.15% of individuals have an operation listed. There are 626,728 total diagnoses and 224,007 total operations. The relative frequency of each of these diagnostic clinical codes is shown in Figure 3.5 and the frequency of each of these operation clinical codes is shown in Figure 3.6, where data is presented in a similar way to [48] with the red dashed line representing the mean frequency.

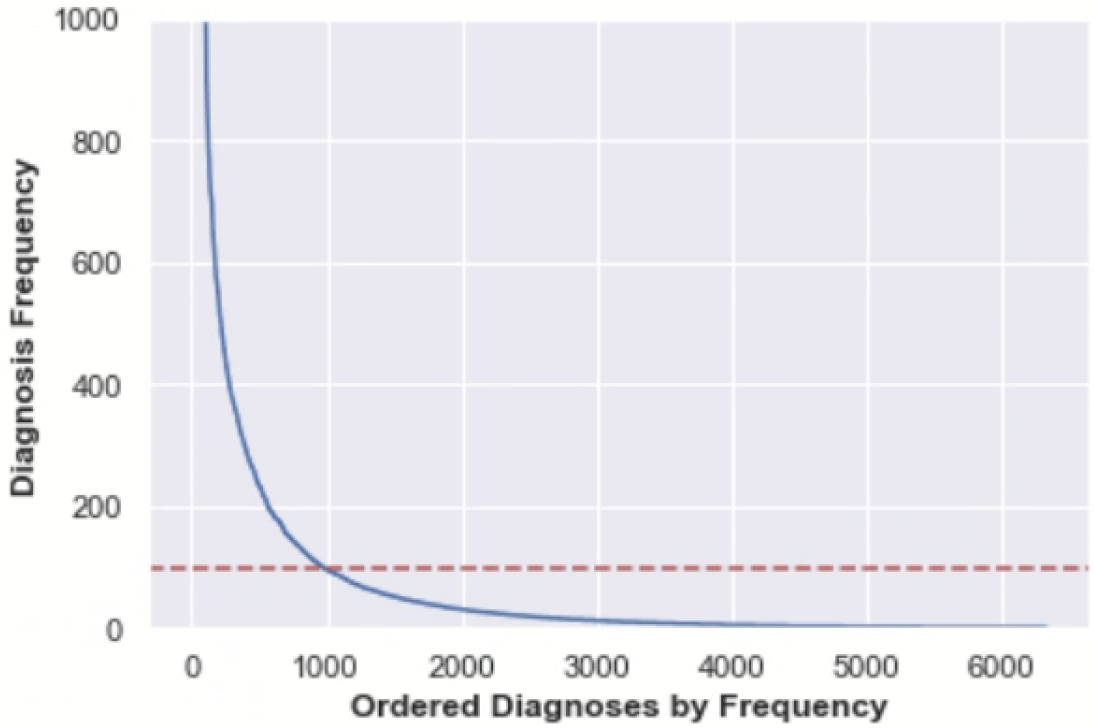


Fig. 3.5 Distribution of Operation Occurrence in Cohort

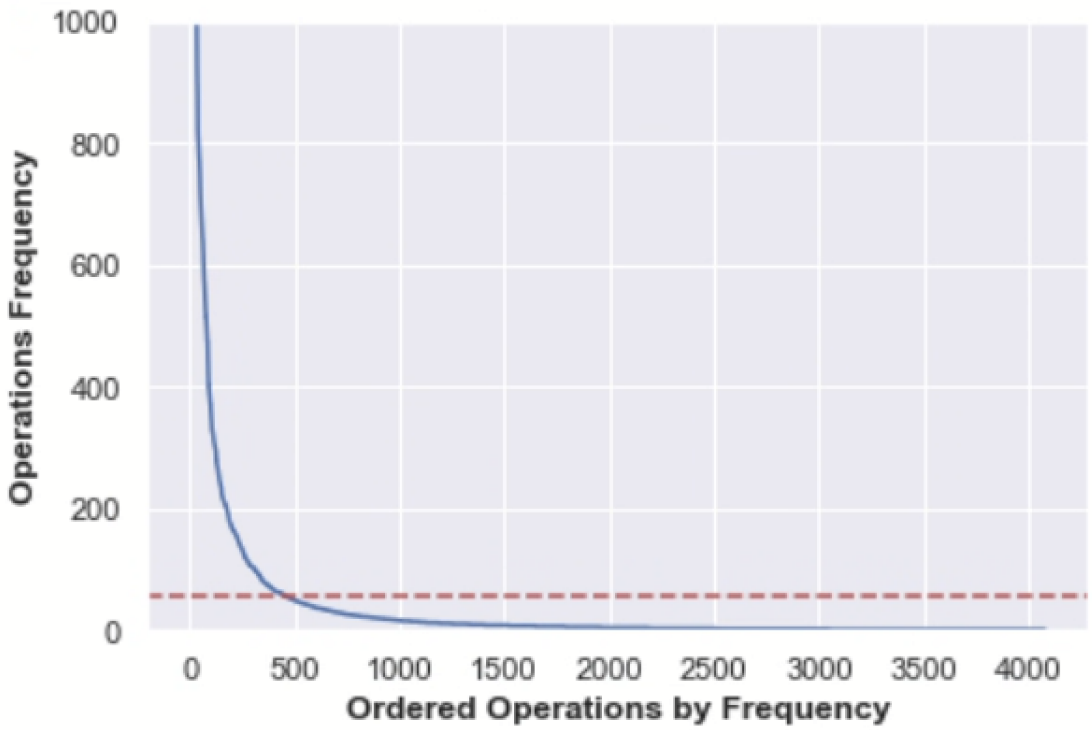


Fig. 3.6 Distribution of Diagnosis Occurrence in Cohort

As shown in Figure 3.5 and Figure 3.6, the distribution of diagnoses and operations are both heavily skewed. Most diagnoses and operations in the data are exceedingly rare while a small number are quite common. This disparity in data frequency will be addressed in the following chapter. Since the diagnosis data includes mental health diagnoses, this information could be combined with read codes from the GP data to measure mental health diagnoses for children in the cohort. Utilizing this approach could be helpful to both validate the quality of the mental health information within the social care data and gain more granular data on diagnoses for prediction. By using established code lists [22], one could use this diagnosis data for measurement of mental health outcomes for mental health prediction algorithms. Although this approach was not utilized for this thesis, it is currently being explored by other members of Dr. Anna Moore's research team.

3.3.5 Measuring Mental Health Outcomes

The accurate measurement of mental health outcomes is critical to the overall performance and viability of the methods. The health status of individuals within the social care databases is tracked over time including monitoring of their mental health status. In the social care datasets, positive mental health status corresponds to individuals diagnosed with mental

health problems by a medical practitioner or individuals on a waiting list for such services. These diagnoses include mental health problems such as depression, eating disorders, and self-harming. However, the diagnoses do not include learning disabilities or substance misuse if they are not accompanied by other mental health issues. Utilizing the data provided from the social care datasets, I measured the mental health status of the cohort. Results are shown in Table 3.6.

Table 3.6 Mental Health Status of SAIL Cohort

Mental Health Status	**# of Individuals**	**Percent of Cohort**
Positive	6,706	14.36
Negative	39,998	85.64
Total	**46,704**	**100**

Overall, we can see in Table 3.6 that the vast majority of individuals within the dataset (85.64%) do not have a diagnosed mental health problem. Hence, the dataset is unbalanced since there are many more healthy children than children with mental health issues; this issue of class imbalance will be tackled in later sections. The class imbalance in the dataset is likely exacerbated by the underdiagnosis of mental health issues within children since many cases of children experiencing mental illness go undetected [21]. Children who experience barriers to receiving proper psychiatric support may be especially likely to have a missed diagnosis.

Given the binary nature of the mental health labels, I frame the prediction problem as a classification problem discriminating between the presence/absence of mental health problems rather than classifying more specific mental health pathology. I therefore will frame this predication task as one of binary classification. This choice to focus on binary classification of mental health is supported by the literature. Interventions relating to child mental health are often similar regardless of the underlying pathology. Moreover, prior research has shown that the labels of a psychiatry diagnosis are not sufficient to create high-performing models due to the fact that psychiatric diseases are often heterogeneous, multifactorial, and highly comorbid [40, 29]. Together, these factors make psychiatric diseases particularly difficult to disentangle. These diagnostic challenges can be magnified for children. Various approaches to dealing with the heterogeneous nature of mental health labels exist such as predicting symptoms or functional consequences [29]. However, this work will focus upon the presence/absence of any mental health disorder as our primary

outcome measure. Although a degree of granularity is lost with this approach, it provides the best modelling approach given the limitations of the data.

There are two classes of tasks related to the prediction of child mental health: diagnosis and prognosis. Diagnosis refers to assessing whether an individual currently is experiencing mental illness. In contrast, prognosis refers to assessing whether an individual will experience mental illness in the future. Due to the low quality of time assessments, the most natural formulation for early identification of child mental health is diagnosis. Further, diagnosis most closely aligns with existing node prediction tasks within the GNN literature. Now that I have described the SAIL database in depth and defined the prediction task, the next chapter will build upon this characterization work by focusing on the development of similarity metrics that utilize the SAIL Databank for graph construction.

Chapter 4

Similarity Metrics for Graph Construction

4.1 Motivation

GNN models provide additional predictive power by incorporating information from a node's neighbors into the feature representation for that node. In our prediction task, we intend to create a neighborhood that informs the prediction of child mental health. Thus, the overall goal of the creation of these similarity metrics is to group together individuals who share pertinent characteristics that may be predictive of their mental health status. Hence, by exploring various similarity metrics, we aim to create an intelligent graph structure that will provide additional predictive power by creating edges between "similar" patients. This approach can be interpreted as adding sparsity to the graph structure. Although a fully-connected graph could theoretically recreate this structure, the approach proposed here decreases noise and lowers computational overhead during training and evaluation of GNNs. However, one distinct disadvantage of these methods is that they are computationally intensive. This computational overhead is driven by the pairwise computations which have a computational complexity of $O(n^2)$, where n is the number of patients in the dataset.

4.2 Data Preparation for Similarity Metrics Calculation

To prepare the data for similarity metrics calculation, I converted all categorical variables to one-hot encodings. To ensure that continuous and categorical variables have a similar magnitude of impact on similarity metric calculations, continuous variables were mapped to a categorical space by splitting each continuous variable into five bins, each bin containing 20%

of the data. In addition, for both categorical and continuous variables, I represented missing data as its own unique value within the model with a corresponding one-hot encoding.

Information relating to diagnoses and operations was also incorporated into the similarity metrics. Due to the high dimensionality of this data, there is no standard way of incorporating this information into clinical models. In this application, to ensure that the diagnosis information does not dominate the score of the similarity metrics, I implemented a prevalence cut-off of 5% for all 2-character codes. I chose to focus solely on the first two characters of the ICD-10 and operations codes because 2-character codes represent a broad category of related diseases. For example, the ICD-10 code "E0" refers to disorders of the thyroid gland. Although this approach inherently leads to the loss of pertinent disease information, this intermediate level of granularity is likely sufficient for modelling the association between physical health diagnoses and adverse mental health events. Utilizing this approach, I included 34 unique diagnoses and seven unique operations as variables for the similarity metrics. These diagnoses and operations were incorporated into the model using one-hot encodings.

4.3 Baseline Similarity Metric

As a baseline, I first focused on an approach that assigns equal weight to each feature. Given the one-hot encodings described above, the similarity metric can then be defined for patient i and patient j as:

$$M_{i,j} = \sum_{f \in Features} 1[f_i = f_j] \tag{4.1}$$

The similarity score can then be normalized by dividing by the total number of features to yield a score that varies from 0 to 1. This normalized metric is defined such that $M_{i,i} = 1$ and such that two patients who share no features in common have a similarity score of 0. Although the simplifying assumption of equal importance of features is limiting, this approach is unsupervised, satisfies the principle of parsimony, and provides the additional benefit of not requiring the use of any learnable parameters.

However, this method is quite limited in its expressive power. For instance, this method treats all features independently and does not account for associations between features. Furthermore, closely tied to the No Free Lunch Theorem in ML, this approach directly incorporates the patient data first into the formation of the graph and later into the node features of the graph. Hence, there is a substantial redundancy of model information that could limit the viability of this approach. Figure 4.1 shows an assessment of whether the top

five closest individuals based upon this similarity metric are more likely to have the same mental health outcome for all individuals within the validation set.

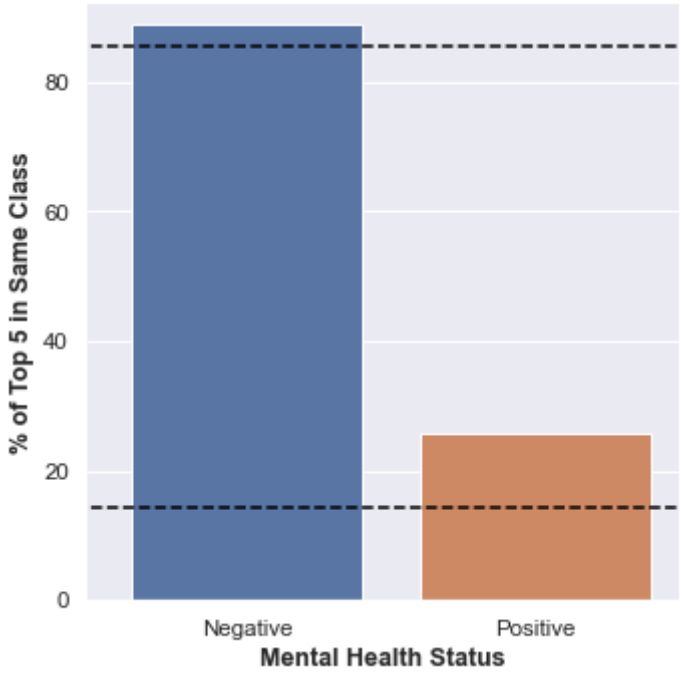


Fig. 4.1 Discriminative Performance of Baseline Distance Metric

In Figure 4.1, the top dashed black line corresponds to the percentage of individuals without a mental health diagnosis in the cohort (85.64%), while the bottom dashed black line corresponds to the percentage of individuals with a mental health diagnosis (14.36%). In this and subsequent analyses, error bars are not included due to computational constraints. One would expect a similarity metric that has no predictive power to have performance roughly corresponding to these dashed lines based upon random chance alone. Model performance for the validation dataset is above the dashed lines indicating that the metrics are performing above the level of null performance. Individuals positive for mental health on average have 25.62% of their five closest neighbors also positive for mental health in the validation cohort. Overall, these results indicate that this baseline similarity metric provides utility in helping to discriminate between individuals with and without a mental health diagnosis. The results are nearly identical for both the training set and validation set, an unsurprising result given that this approach does not contain any parameters that are trained using the training data. Since the following metrics will all be trained on the training data, performance will only be shown on the validation cohort.

4.4 Full Diagnosis Graph Approach

This section builds on, and follows the notation of, the work previously done by Rocheteau et al. [48]. ICD-10 diagnoses are first transformed into a matrix of size m by N where m represents the number of unique diagnoses and N represents the number of patients. To maintain the hierarchical structure, diagnoses are included at each class level. If the specific code is "G11.2", then "G", "G1", "G11", and "G11.2" will all be included. A prevalence cut-off of .5% is applied, corresponding to keeping 673 out of the 7,613 total diagnoses. Since diagnoses are hierarchical, if a diagnosis does not meet the prevalence threshold, it will still be included via any parent classes that meet the threshold. The diagnostic similarity score $M^D_{i,j}$ of two patients i and j can be defined as:

$$M^D_{i,j} = a \sum_{\mu=1}^{m} (D_{i\mu} D_{j\mu} (d_\mu^{-1} + c)) - \sum_{\mu=1}^{m} (D_{i\mu} D_{j\mu}) \tag{4.2}$$

In Equation 4.2, a and c are tunable constants and d_μ refers to the frequency of a specific diagnosis. This equation corresponds to assigning greater similarity to patients who share more diagnoses, while upweighting the importance of sharing a rare diagnosis and downweighting individuals with many diagnoses.

In the original paper detailing this similarity metric method [48], they solely looked at diagnoses. Given that the SAIL dataset also contains operation codes, I conjecture that there could be additional predictive power from also incorporating in information regarding operations. Applying the same prevalence cut-off, we keep 74 out of the 3,174 total operations. Then, the operations similarity score $M^O_{i,j}$ of two patients i and j can be defined as:

$$M^O_{i,j} = a \sum_{\mu=1}^{m} (O_{i\mu} O_{j\mu} (o_\mu^{-1} + c)) - \sum_{\mu=1}^{m} (O_{i\mu} O_{j\mu}) \tag{4.3}$$

To ensure that the two similarity scores are in a form amenable to combination, each diagnostic and operation raw score is converted to a percentile. To account for the different scales of the metrics, the percentiles are computed for operations and diagnoses separately. Then, the similarity scores can be combined using Equation 4.4.

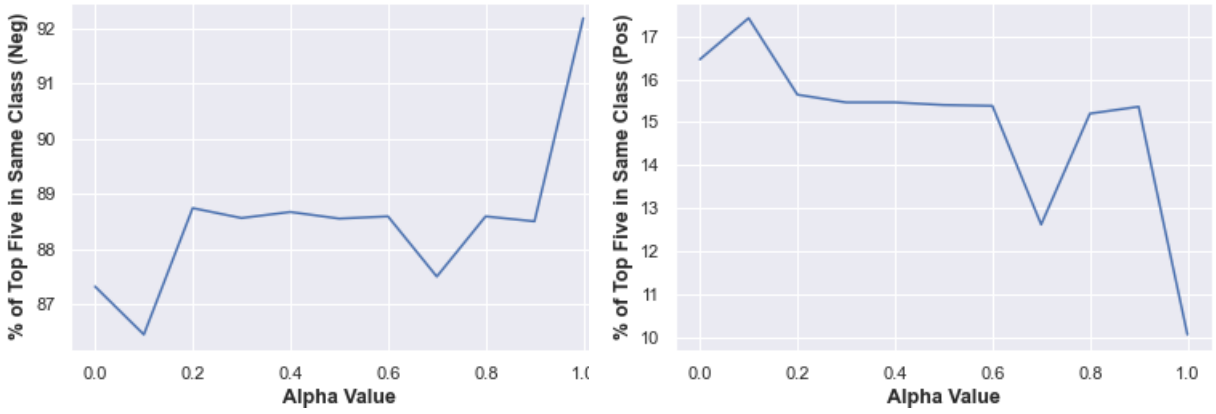
$$M_{i,j} = \alpha * M^O_{i,j} + (1 - \alpha) * M^D_{i,j} \tag{4.4}$$

By using α, we can assess the performance of a metric incorporating information from both the diagnoses and operations, while weighting which information receives more attention.

An α value of 0 would correspond to only considering diagnosis information, an α value of .5 would correspond to equal weighting, and an α value of 1.0 would correspond to only considering operations information. Results from assessing this metric on the validation set can be seen in Table 4.1 and Figure 4.2.

Table 4.1 Discriminative Performance of Full Diagnosis/Operation Similarity Metric on Validation Set

α Value	% Same Neg	% Same Pos
0	87.32	16.46
.1	**86.45**	**17.42**
.2	88.74	15.64
.3	88.56	15.46
.4	88.67	15.46
.5	88.55	15.40
.6	88.59	15.38
.7	87.50	12.62
.8	88.59	15.2
.9	88.50	15.36
1.0	92.18	10.06



(a) Performance for individuals with mental health diagnosis

(b) Performance for individuals without mental health diagnosis

Fig. 4.2 Discriminative Performance of Full Diagnosis/Operations Graph as Function of Weighting Term α

In Table 4.1 and and Figure 4.2, the model tends to perform better when giving more weight to the diagnosis information than to the operations information. This aligns with our intuition because individuals in the dataset tend to have many more diagnoses than operations. In fact, more than 50% of individuals have no operations information at all. In this and subsequent analyses, I will focus predominantly upon the ability of the distance metric to cluster individuals with a mental health diagnosis rather than the ability to cluster individuals without a mental health diagnosis. I choose this focus because there is substantially more variance in the performance of clustering of individuals with a mental health diagnosis and because clustering individuals with a mental health diagnosis will be more important for the future prediction task. The model achieves the best performance with an α value of .1, corresponding to weighting the diagnosis information as nine times more important. However, even for this optimal setting of α, the performance for both classes is only marginally better than chance.

It is important to note that this approach cannot be directly compared to the other approaches as it does not take into account all of the variable information; rather, this framework attempts to insert additional information into the model by intelligently incorporating information from rare diagnoses. Nevertheless, the metric does not contribute much predictive capability. This finding stands in stark contrast to the performance of this general approach in the context of length of stay prediction in the Intensive Care Unit (ICU) where the metric greatly improves model performance. This discrepancy can likely be attributed to the nuanced relationship between diagnosis information and child mental health prediction [6]. For instance, individuals with a chronic health issue are more likely to experience depression [5]. Additionally, children oftentimes accumulate physical health diagnoses on their way to an ultimate mental health diagnoses [14]. Hence, physical and mental health are inextricably linked, and the relationships between the two may be difficult for this metric to effectively untangle in this dataset. In contrast, the relationship between diagnostic history and length of stay in an ICU may be more easily modelled by this approach.

4.5 Standard Autoencoders

Autoencoders are a family of artificial neural networks that learn efficient codings of data in an unsupervised fashion. They consist of an encoder which maps the input data through a series of layers to a latent space and a decoder which then attempts to regenerate the input from the representation in the latent space [25]. The motivation underlying the application of autoencoders to this task stems from the inability of simple similarity metrics like those

described above to model associations between features. Thus, by mapping to a reduced latent space, I hope to model the underlying associations between different features in order to create a similarity metric that more accurately separates individuals based upon their mental health status.

The form of the autoencoder applied to this problem follows a simple structure. It consists of an encoding layer comprised of a singular hidden layer with ReLU activations, a bottleneck layer (called the latent layer in this thesis), and a decoding layer that mirrors the structure of the encoding layer to map back to the input space. This particular autoencoder structure can be seen in Figure 4.3 using a graphic created with NN-SVG.

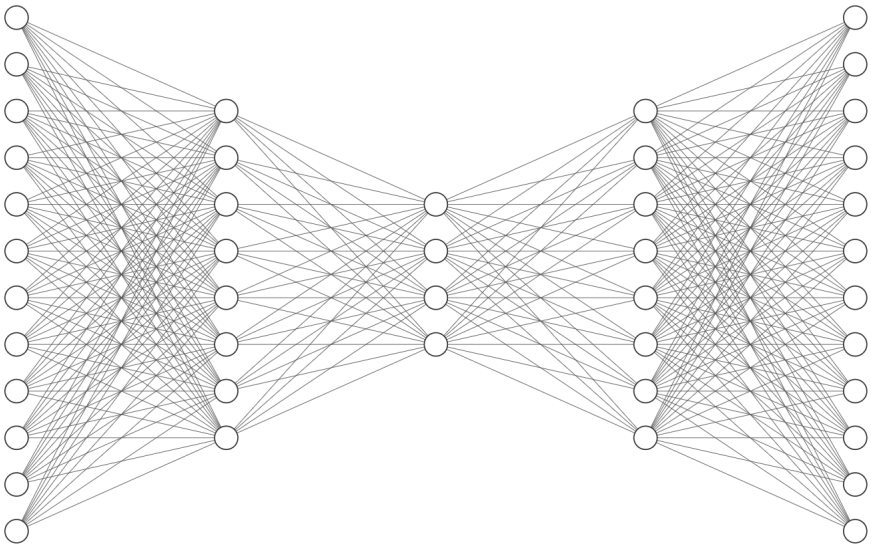


Fig. 4.3 Autoencoder Structure (Number of Nodes Not to Scale)

The model utilizes an L2 loss function over the latent space, as defined by Equation 4.5 where X is the original input and $\hat{X}$ is the reconstructed input created by the autoencoder.

$$\text{Reconstruction Loss} = (X - \hat{X})^2 \tag{4.5}$$

To use the autoencoder as a similarity metric, the autoencoder model is first trained to effectively reconstruct the data by training on the training dataset. After training, the encoding portion of the autoencoder network can be used to create latent space embeddings for each patient. Euclidean distance is then utilized to measure distance between vectors in this latent space. In this framework, increased similarity scales with decreased Euclidean distance in the latent space. Performance of the autoencoder was assessed using various sizes

for both the hidden layer the latent layer. The results on the validation dataset are shown in Table 4.2 and Figures 4.4-4.5.

Table 4.2 Discriminative Performance of Autoencoder Similarity Metric with Networks of Varying Size

Latent Layer Size	**Hidden Layer Size**	**% Same Neg**	**% Same Pos**
10	25	88.48	24.52
10	50	88.43	22.46
10	75	88.54	22.88
10	100	88.10	21.98
15	25	88.21	22.02
15	**50**	**89.14**	**26.16**
15	75	88.34	22.44
15	100	87.87	22.22
20	25	88.79	23.60
20	50	88.30	24.70
20	75	88.32	22.66
20	100	88.52	23.00
25	25	88.73	24.14
25	50	88.35	23.32
25	75	88.45	23.9
25	100	88.91	24.38

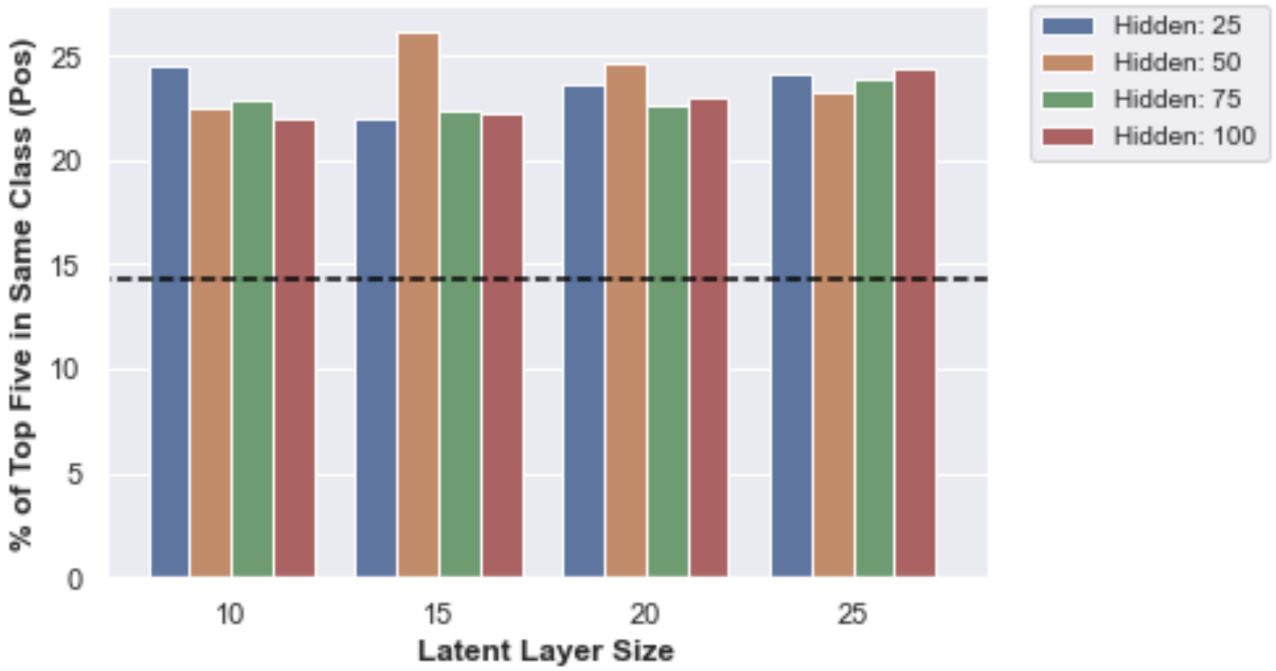


Fig. 4.4 Discriminative Performance of Autoencoder Similarity Metric with Networks of Varying Size (Individuals with Mental Health Diagnosis)

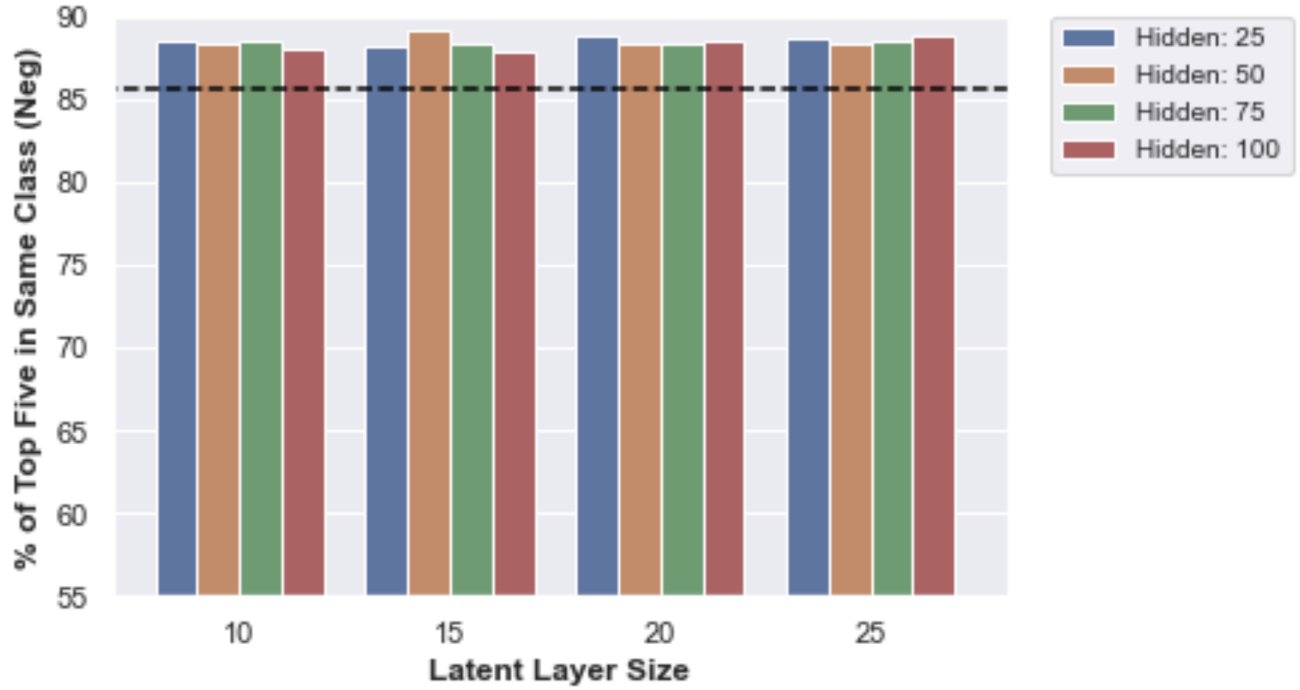


Fig. 4.5 Discriminative Performance of Autoencoder Similarity Metric with Networks of Varying Size (Individuals without Mental Health Diagnosis)

Examining Table 4.2 and Figures 4.4-4.5, all models achieve relatively similar results which are on par with the results from the baseline approach. There are not any easily discernible trends relating to the impact of the size of layers on model performance. The model with a latent layer size of 15 and a hidden layer size of 50 appears to achieve the best performance. 89.14% of the top five closest neighbors for an individual without a mental health diagnosis do not have a mental health diagnosis, and 26.16% of the closest neighbors for an individual with a mental health diagnosis also have a mental health diagnosis. This analysis could be expanded by also exploring how altering the number of decoding layers and encoding layers could affect model performance. It is possible that deeper networks could improve model performance. Nonetheless, the results as presented indicate that distance between encoded points in this latent space is a sensible metric.

It is important to note the trade-off inherent in choosing the dimension of the latent space. In typical autoencoder models, smaller latent spaces protect against overfitting by forcing the model to learn an effective lower-dimension representation, while larger latent spaces provide additional modelling flexibility that allows for more effective reconstruction of training data points. In our specific case, there exists a trade-off between how faithful the reconstruction is to our original data points and the ability to learn a useful dimensionally-reduced latent representation.

4.6 Dual-Loss Autoencoders

Experiments from the previous section validate the approach of using autoencoders to calculate distances between patients in a reduced latent space. Expanding upon this idea, I assessed how an additional loss function could further induce the latent space to learn a representation helpful for classifying individuals. An output layer with a single node is connected to the latent bottleneck layer, while the rest of the model framework including the decoding layer is left unaltered. This output layer is utilized to predict the mental health status of an individual. The structure of this additional use of the latent layer can be seen in the diagram in Figure 4.6, which is created using NN-SVG. The strucuture includes a sigmoid activation on the output layer and the use of binary cross entropy loss with a weighted loss function (described in detail in Chapter 5) which is not propagated through the decoding layer.

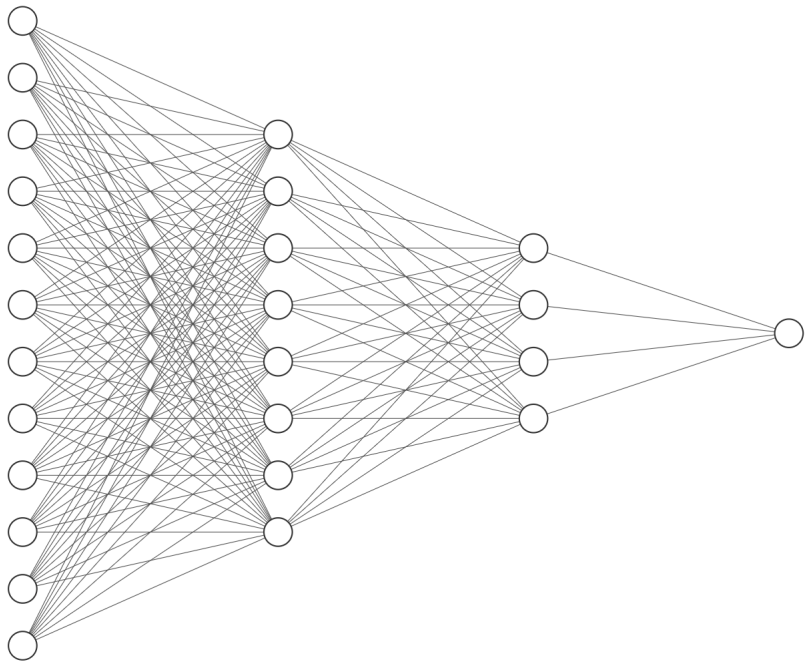


Fig. 4.6 Partial Autoencoder Structure for Dual-Loss (Number of Nodes Not to Scale)

The rationale underlying this dual-loss structure is to learn a latent space structure that is useful for both reconstruction of the original input and prediction of an individual's mental health status. Balancing the training of these two different loss functions can be achieved by introducing a weighting term as described in Equation 4.6.

$$\text{Loss Total} = \alpha * (\text{Reconstruction Loss}) + (1 - \alpha) * (\text{Prediction Loss}) \tag{4.6}$$

The autoencoder model is trained on the training data with the two loss functions. Then, the encoding portion of the autoencoder network is used to create latent space embeddings for each of the training points with Euclidean distance utilized to measure similarity in this latent space. Model performance using various latent layer sizes and values for α is shown in Table 4.3 and Figures 4.7-4.8. All analyses are conducted with a hidden layer size of 100 for both the encoding layer and decoding layer, and performance is assessed on the validation dataset.

Table 4.3 Discriminative Performance of Dual-Loss Autoencoder Similarity Metric with Varying Latent Layer Size and Varying α

Latent Layer Size	$\boldsymbol{\alpha}$	**% Same Neg**	**% Same Pos**
10	.9	87.54	23.46
10	.75	88.25	23.18
10	.5	88.47	24.92
10	.25	89.52	28.72
10	.1	89.77	29.90
10	.01	89.45	30.94
15	.9	88.45	23.38
15	.75	88.22	25.56
15	.5	88.87	24.50
15	.25	89.51	28.70
15	.1	89.59	29.98
15	**.01**	**89.35**	**32.38**
20	.9	89.18	25.64
20	.75	88.81	24.02
20	.5	89.27	26.22
20	.25	89.36	29.36
20	.1	89.88	29.84
20	.01	89.55	31.96

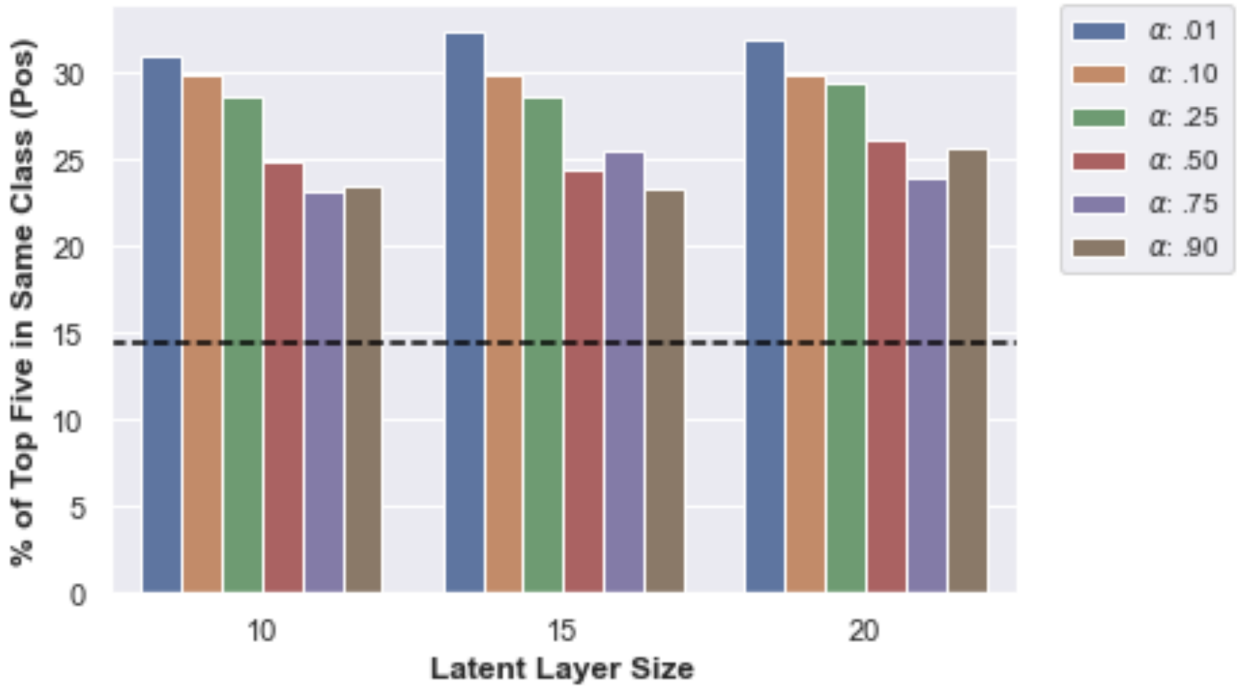


Fig. 4.7 Discriminative Performance of Dual-Loss Autoencoder Similarity Metric with Varying Latent Layer Size and Varying α (Individuals with Mental Health Diagnosis)

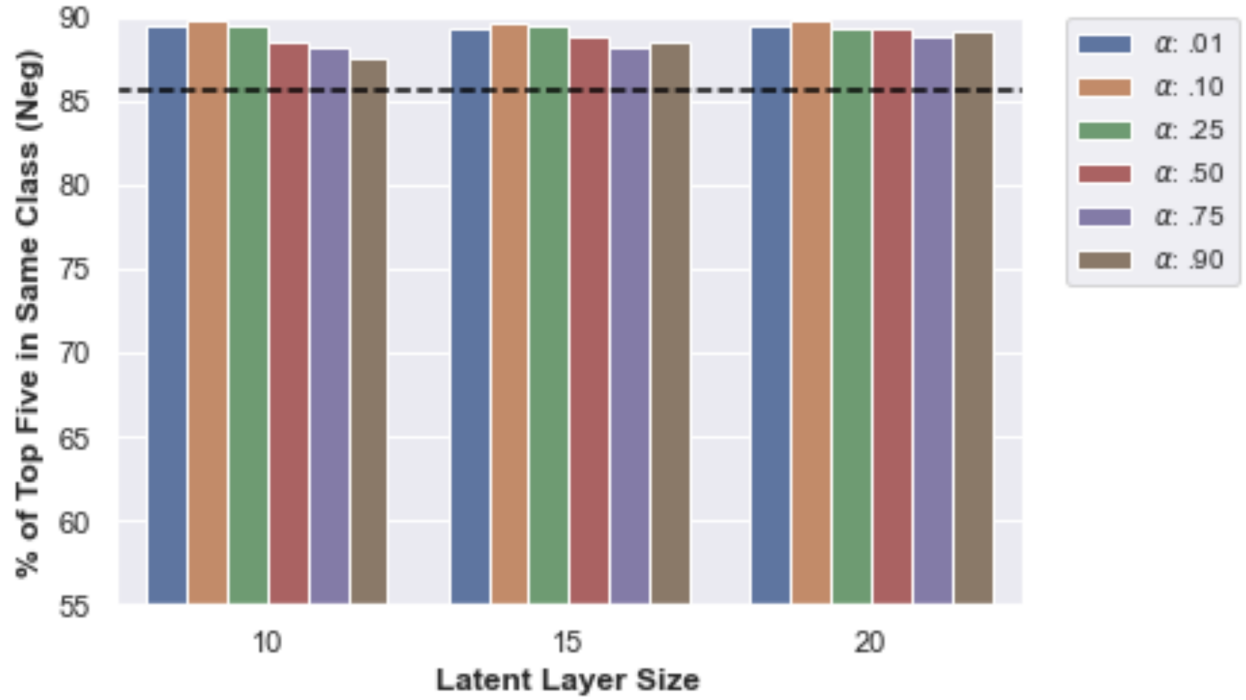


Fig. 4.8 Discriminative Performance of Dual-Loss Autoencoder Similarity Metric with Varying Latent Layer Size and Varying α (Individuals without Mental Health Diagnosis)

In Table 4.3 and Figures 4.7-4.8, the model with the best performance has a latent layer with 15 nodes and uses an α of .01, corresponding to weighting the prediction loss as 99 times more important than the reconstruction loss. Overall, models with lower values of α typically perform superior to the models with higher values of α. This weighting trend is partially explained by the reconstruction loss typically being an order of magnitude larger than the prediction loss for this specific task. This difference in magnitude cannot fully account for the effect, indicating the importance of the prediction loss in identifying an effective latent space. The training dynamics for the best-performing metric with 15 nodes in the latent layer and an α of .01 can be seen in Figure 4.9.

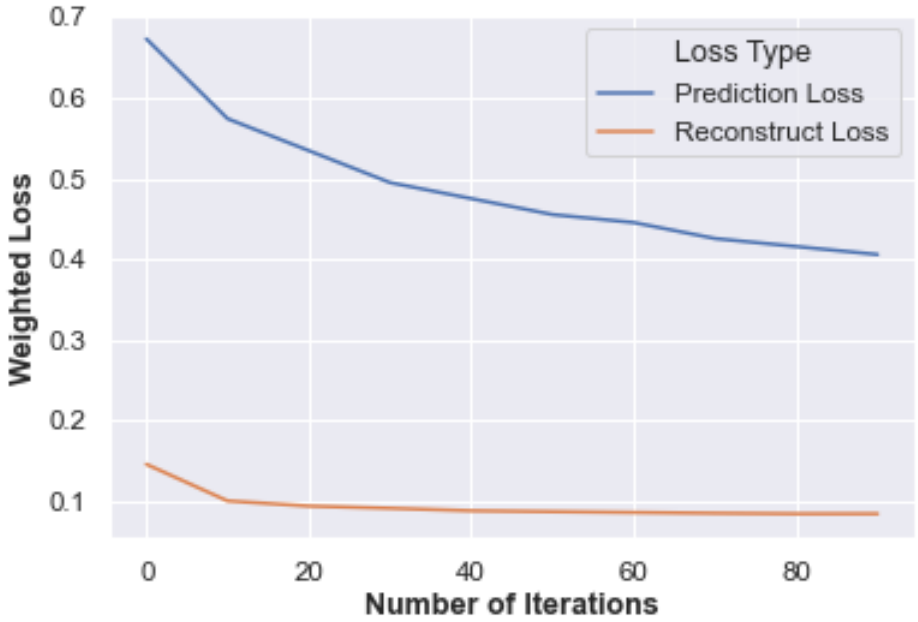


Fig. 4.9 Training Dynamics Dual-Loss Autoencoder (Loss Weighted by Alpha Value)

In Figure 4.9, we can see that the model is able to drive down both the reconstruction loss and the prediction loss, effectively creating a latent space that is well-suited for both tasks. Although the model focuses predominantly on driving down the prediction loss, the model is able to effectively achieve solid performance on both tasks. Finally, in order to better understand the dynamics of the latent space created by the autoencoder, t-SNE with two dimensions was used to visualize the latent space for the best-performing Dual-Loss autoencoder for two different random seeds. Results are shown in Figures 4.10- 4.11.

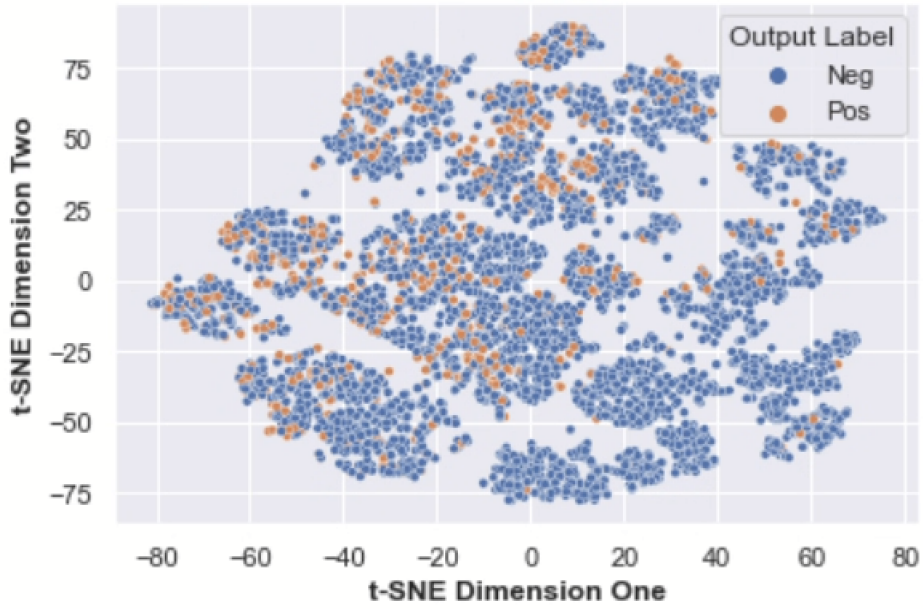


Fig. 4.10 2-Dimensional t-SNE Visualization of Latent Space of Best-Performing Dual-Loss Autoencoder 1st Seed

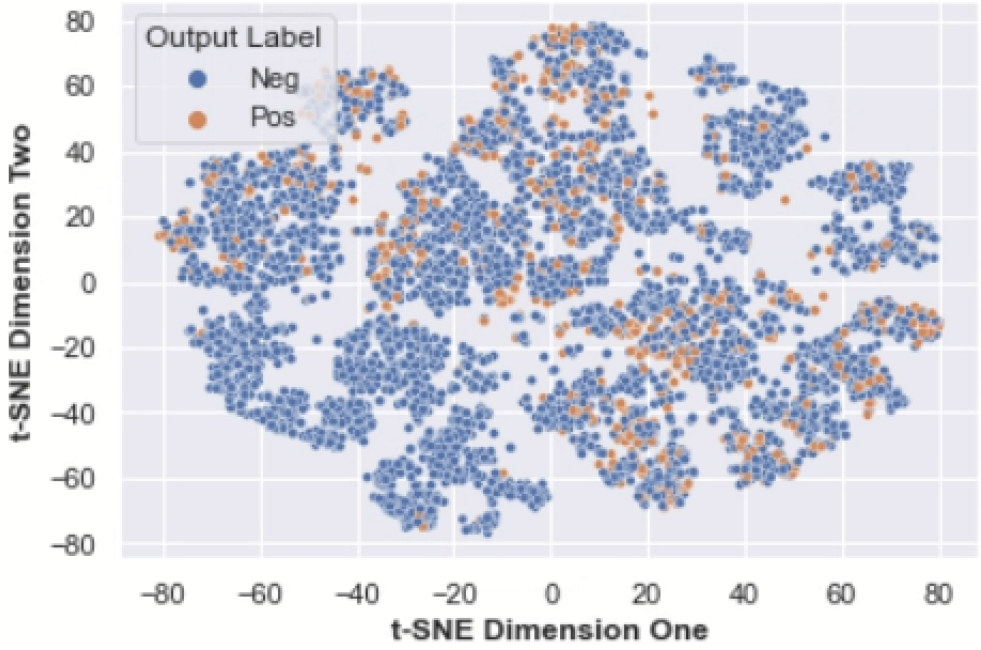
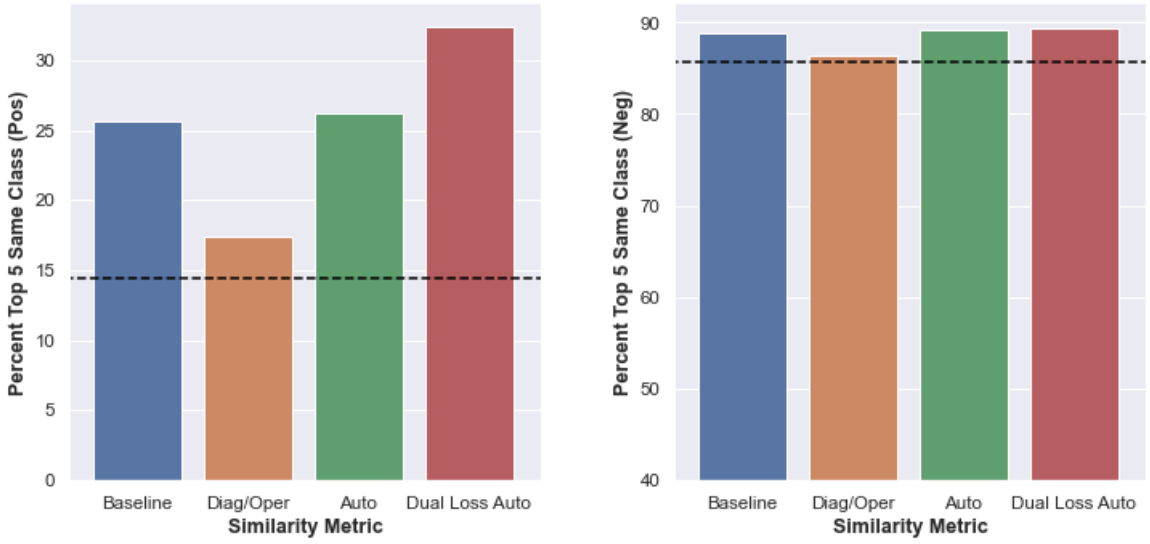


Fig. 4.11 2-Dimensional t-SNE Visualization of Latent Space of Best-Performing Dual-Loss Autoencoder 2nd Seed

Figures 4.10- 4.11 demonstrate that the latent space is able to effectively cluster individuals with a mental health diagnosis. In this two-dimensional space, individuals with a mental health diagnosis are often surrounded by other individuals with a mental health diagnosis in various hubs dispersed throughout the graph. Moreover, in the bottom right of Figure 4.10 and the bottom left of 4.11, there are regions of the graph dominated by individuals without mental health diagnoses. Overall, these results vindicate the choice of dual-loss autoencoder for its ability to induce a latent space that separates individuals with and without a mental health diagnosis. Conducting a more thorough interpretation of the t-SNE space could prove informative. For instance, one could assess which characteristics of individuals correspond to areas dominated by individuals with high risk of mental health disorders.

4.7 Comparing Similarity Metrics

To assess the relative advantages of each technique, the best performing models from each of the different approaches are collated and compared in Figure 4.12.



(a) Performance for individuals with mental health diagnosis

(b) Performance for individuals without mental health diagnosis

Fig. 4.12 Discriminative Performance Comparison of Best-Performing Similarity Metrics

Figure 4.12b demonstrates the all four models perform comparably for individuals without a mental health diagnosis, though the diagnostic graph approach shows slightly worse performance. However, for individuals with a mental health diagnosis, the dual-loss autoencoder demonstrates superior performance where individuals with a positive mental

health diagnosis are, on average, grouped with 31.58% other individuals with a mental health diagnosis, more than two times greater than the expected value based upon the mental health prevalence of 14.36% in the dataset. Both the baseline and single-loss autoencoder approaches demonstrate solid performance as well. In order to assess how the inclusion of these metrics for graph construction affects GNN model performance, the baseline metric and top-performing dual-loss autoencoder will be utilized for graph construction for GNNs in the following chapter.

Chapter 5

Mental Health Prediction With Graph Neural Network Models

5.1 Motivation

Having created similarity metrics that help to cluster individuals, I will now evaluate the performance of these metrics. This assessment will focus on two similarity metrics (dual-loss autoencoder and baseline similarity metric) and three GNN layer types (SAGE, GAT, and MPNN). Performance of GNNs will be compared to a baseline NN model, and the potential of the derived models to exacerbate disparities in care will be explored by conducting analyses of model fairness.

5.2 Data Preparation

Data was prepared following the same procedure for the similarity metrics with minor changes reflecting the difference in the tasks. To retain the most information from continuous variables while normalizing data magnitudes, continuous variables were standardized using sample means and standard deviations with absolute cut-offs applied at ± 4 standard deviations from the mean. To retain as much useful clinical information as possible, a prevalence cut-off of 2.5% was used for diagnosis and operation codes, leading to the inclusion of 61 unique diagnoses and 26 unique operations. The similarity/distance metrics were passed as edge attributes for MPNN models.

In the dataset, there are many more healthy individuals than individuals with a mental health diagnosis; thus, accuracy is a misleading metric as a model could trivially obtain high accuracy by predicting all children to be healthy. To address this problem, the weight of the

training samples are adjusted to obtain a classifier with real-world utility. The formula for weighting is shown in Equation 5.1 for individuals with a mental health diagnosis and shown in Equation 5.2 for individuals without a mental health diagnosis.

$$\text{Positive Weight} = \frac{\text{Number of Samples}}{2 * (\text{Number of MH + Individuals})} \tag{5.1}$$

$$\text{Negative Weight} = \frac{\text{Number of Samples}}{2 * (\text{Number of MH - Individuals})} \tag{5.2}$$

There is a prevalence of mental health diagnoses of 14.36% in this dataset, corresponding to a positive class weight of 3.48 and a negative class weight of 0.58. This is equivalent to upweighting the loss of the positive samples by approximately six times. To account for the trade-off between performance on the positive samples and negative samples, AUROC is utilized as the primary assessment of model performance. AUROC can be interpreted as representing the probability that a classifier will rank a randomly chosen positive instance higher than that of a randomly chosen negative instance [12]. For cases of extreme class imbalance where there is low value in true negative decisions, AUPRC has been proposed as an effective measure. Since class imbalance is not incredibly large and there is substantial value in true negative decisions (correctly identifying a children without a mental health diagnosis) in this context, I designate AUROC as the primary metric. I use AUPRC as a secondary metric to obtain a more nuanced view of overall model performance. Finding an ideal threshold that balances the clinical impact of false positives and false negatives is left for future work.

5.3 Baseline Models

To my knowledge, no one has attempted to predict child mental health using SAIL data; as such, there are no existing benchmarks against which I can evaluate model performance. Thus, in order to assess the additional predictive power afforded by graphical methods, ML models that treat the variables within SAIL as a flat bag-of-features will be assessed as baselines. Within the psychiatry field, meta-analyses have noted that simpler models can provide comparable performance to more complex models with high levels of interpretability [55, 51]. This is especially true in clinical settings with a low signal to noise ratio [40]. As such, basic Neural Network (NN) models of varying size will be utilized as our baseline for model comparison.

Baseline NN models contain the same features as the node features for the GNN models. The Multilayer Perceptron (MLP) models devised for this baseline task are evaluated with varying sizes and numbers of hidden layers. By assessing model performance for increasingly complex models, I aim to identify the best-performing baseline NN for mental health prediction. Models are implemented using a binary cross entropy loss function with ReLU activations between the layers. Initial results from these models can be seen in Table 5.1 and Figure 5.1.

Table 5.1 Model Performance of MLPs of Varying Configurations

Number of Layers	**Number of Nodes**	**Maximum AUC Validation**
2	20	.786
2	**40**	**.792**
2	60	.789
2	80	.792
2	100	.791
4	20	.786
4	40	.787
4	**60**	**.792**
4	80	.789
4	**100**	**.792**
6	20	.784
6	40	.790
6	60	.788
6	80	.790
6	100	.787
8	20	.622
8	40	.610
8	60	.568
8	80	.541
8	100	.591
10	20	.579
10	40	.593
10	60	.525
10	80	.529
10	100	.500

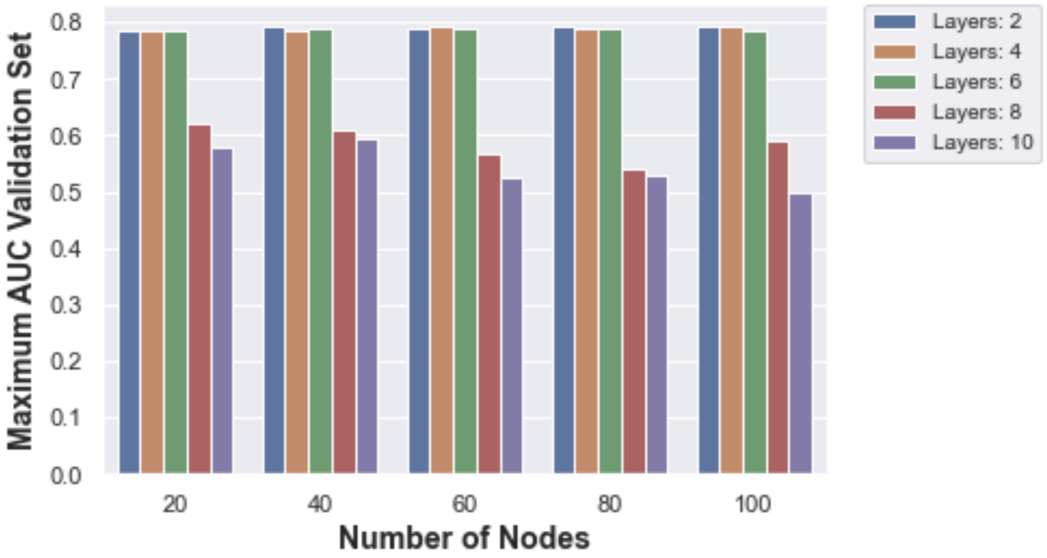


Fig. 5.1 Model Performance of MLPs of Varying Configurations

Table 5.1 visualizes model performance and Figure 5.1 indicates that performance quickly degrades with too may layers, while the number of nodes per layer does not appear to have a substantial impact on model performance. Most network configurations with less than eight layers perform comparably. I chose the overall best-performing model on the validation dataset, a model with four layers and 60 nodes per layer trained for 250 iterations. Training dynamics for the model are shown in Figure 5.2 and results for the model are compared to GNN models in Table 5.2.

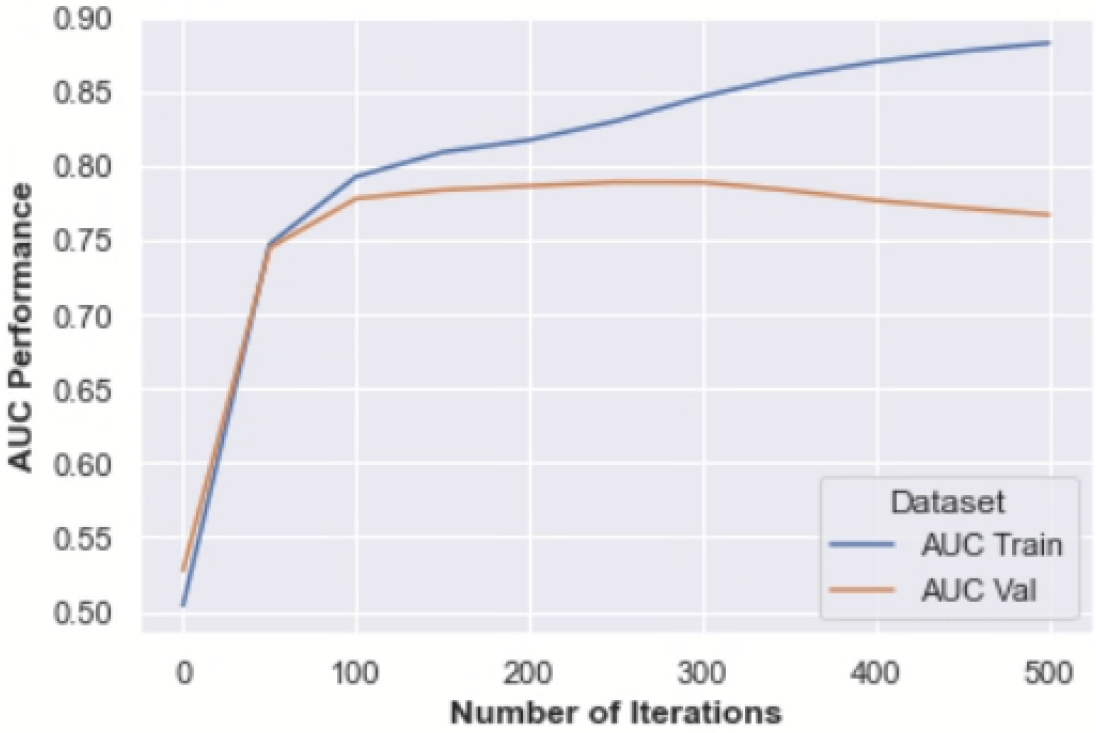


Fig. 5.2 Training Dynamics of Best-Performing Baseline MLP Model with 4 Layers and 60 Nodes Per Layer

The training dynamics shown in Figure 5.2 indicate that the model is able to quickly learn with performance peaking at 250 iterations and the model subsequently overfitting the training data. This overfitting is evident as the training performance continues to increase past 500 iterations, while the validation performance declines after 250 iterations.

5.4 Structure of GNNs

Patient graphs are constructed using the similarity metrics derived and validated in the previous section. These graphical structures consist of patients as nodes with directed edges and edge weights representing similarity relationships between patients. Different methods for graph construction such as establishing a similarity threshold for edge creation could be utilized. However, to ensure that all nodes have neighbors and no nodes dominate as connectivity hubs, each node is connected to itself and the N other closest patients, where N is treated as a hyperparameter. As such, the in-degree of each node is fixed to N, while the out-degree of each node is not restricted. This general graph structure for a single example node is shown in Figure 5.3.

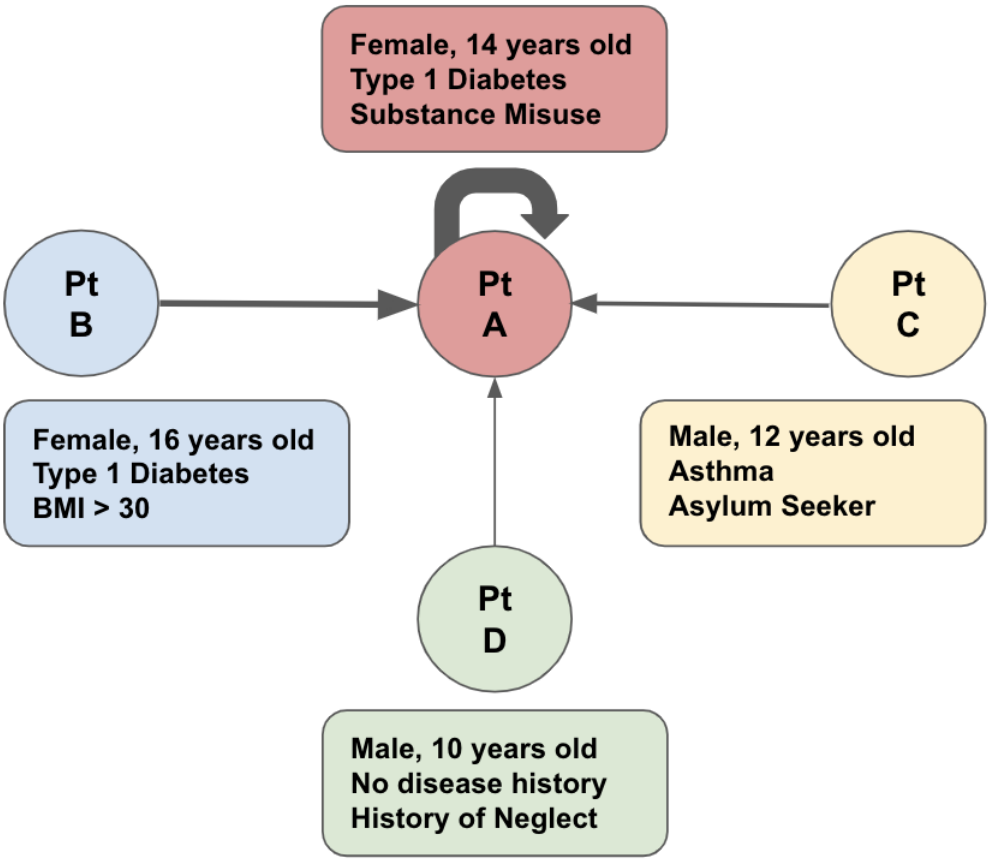


Fig. 5.3 Sample Patient Graph Used for GNN Prediction

In Figure 5.3, patients with higher degrees of similarity have a thicker line connecting them. For instance, Patient A and Patient B are more similar demographically and share a diagnosis of Type 1 Diabetes. This similarity is reflected in the thicker edge connecting them. All of the similarity metrics explored are designed such that patients are more similar to themselves than to any other patient within the dataset.

For this prediction task, an inductive learning approach is undertaken, following the neighborhood sampling techniques proposed by Hamilton et al. [18]. During training, nodes and their neighbors are sampled only from the training set. During testing, nodes are sampled from the test set but neighbors are sampled from both the training set and the test set. The same approach is used to evaluate performance on the validation set. The following GNN layer types are explored: Graph Attention Networks, GraphSAGE, and Message Passing Neural Networks. GraphSAGE and Graph Attention Networks are implemented with two layers. All three models are coded using PyTorch-Geometric.

5.5 Hyperparameter Search and Model Results

To identify optimal model performance settings, an extensive hyperparameter grid search is conducted. This hyperparameter search evaluates models with the two top-performing similarity metrics, varying numbers of neighbors, and different internal model parameters. The best-performing models on the validation set for each layer type and similarity metric are retained. The specific model configurations for top-performing models can be found in the Appendix. Performance is calculated on the validation set and test set with AUROC values and AUPRC values averaged over 15 runs, and results are reported in Table 5.2 with 95% confidence intervals for error margins. Figure 5.4 compares AUROC results on the test set and Figure 5.5 compares AUPRC results on the test set.

Table 5.2 Performance Comparison of GNN Models

Model	**Sim Metric**	**AUROC Val**	**AUPRC Val**	**AUROC Test**	**AUPRC Test**
NN	NA	.781 ± 5.7e-4	.409 ± 1.9e-3	.800 ± 1.1e-3	.441 ± 2.1e-3
SAGE	**Baseline**	**.802 ± 6.9e-4**	**.438 ± 1.5e-3**	**.815 ± 7.0e-4**	**.481 ± 1.4e-3**
SAGE	Dual-Loss Auto	.778 ± 1.1e-3	.386 ± 2.3e-3	.797 ± 9.0e-4	.413 ± 3.9e-3
GAT	Baseline	.773 ± 3.2e-4	.382 ± 2.4e-4	.783 ± 9.1e-4	.427 ± 2.0e-3
GAT	Dual-Loss Auto	.769 ± 6.4e-4	.369 ± 6.4e-4	.780 ± 4.1e-4	.380 ± 1.6e-3
MPNN	Baseline	.790 ± 1.2e-3	.425 ± 1.9e-3	.807 ± 6.9e-4	.467 ± 2.1e-3
MPNN	Dual-Loss Auto	.762 ± 2.1e-3	.361 ± 4.4e-3	.773 ± 1.7e-3	.380 ± 6.3e-3

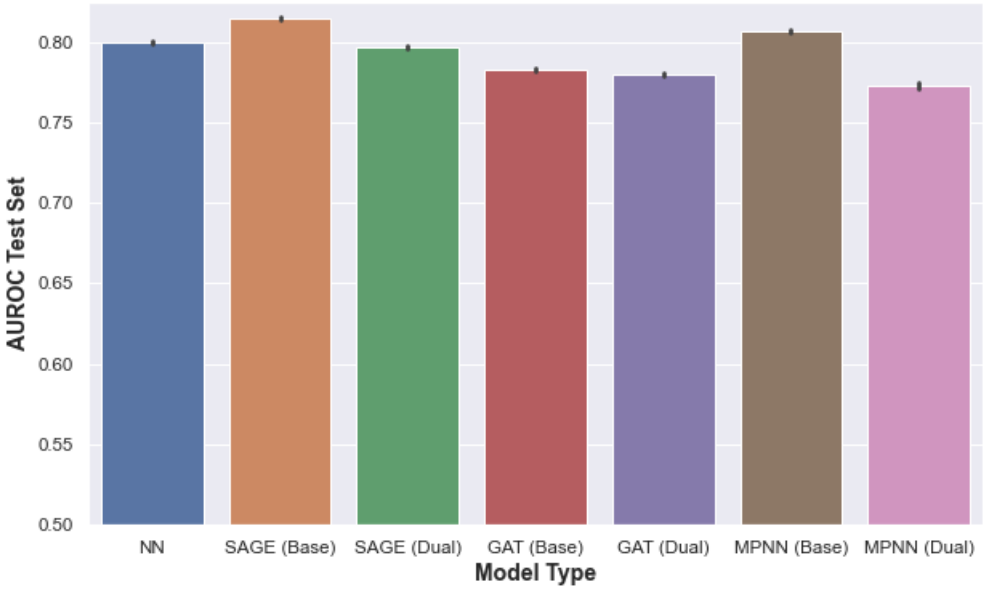


Fig. 5.4 Comparison of AUROC Model Performance on Test Set

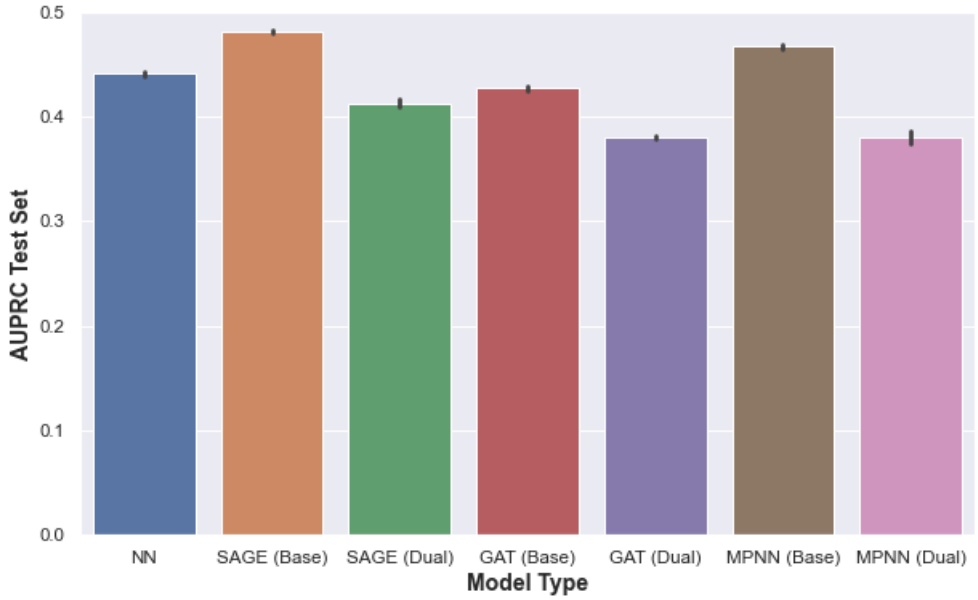


Fig. 5.5 Comparison of AUPRC Model Performance on Test Set

The best-performing model in Table 5.2 on the test dataset is the SAGE model using the baseline similarity metric for graph construction, which achieves an AUROC of .815, a modest but statistically significant improvement over the baseline NN. It is also worth noting the good performance of MPNN and SAGE configurations with both of these models achieving superior performance in comparison to the baseline NN model. In contrast, the GAT models perform slightly worse than the baseline NN model, indicating that this model type may not be well-suited to this particular task. It is possible that the GAT models are not able to effectively apportion the attention between neighbors, leading to overfitting and inferior model performance.

For all GNN models, performance is improved using the baseline autoencoder for patient graph construction, though the extent of improvement varies based upon layer type. This is a surprising result since the dual-loss autoencoder demonstrated superior discriminative performance compared to the baseline metric in Chapter 4. It is possible that there is a more complicated relationship than expected between effectively separating individuals into groups based upon output label and improving GNN performance. One initial conjecture is that the problem stems from the α value of .01 for the dual-loss autoencoder. Since this alpha value corresponds to prioritizing the prediction loss over reconstruction loss, it is possible that individuals in this latent space may be similar in terms of mental health status but starkly different in terms of their input features. Since individuals in a neighborhood may not share many risk factors, GNN model performance may be affected. Further experiments are necessary to disentangle this relationship.

It is also interesting to note that all models performed better on the test set than on the validation set, despite being tuned for performance on the validation set. It is likely that the overall trends relating risk factors and mental health outcomes observed in the test set are more similar to those in the training set than the trends in the validation set. That is, due to chance and the relatively small size of the datasets, the test set tends to be "easier." Finally, scores from AUROC generally aligned quite well with AUPRC scores, and the SAGE model with baseline similarity metric achieved the top scores for both metrics. For this specific task, both metrics appear similarly suited to assess model performance. In relation to clinical utility, the performance of the best-performing model with an AUROC of .815 is promising. A model that performs at this level could merit usage as a predictive tool for measuring child mental health and staging interventions for children at high risk. However, external validation of the model and determination of an optimal model threshold would be necessary before clinical application of such a model should be pursued. Collectively, the results indicate

that infusing additional information into the model via a graph structure provides additional predictive benefit.

5.6 Model Fairness

Algorithmic fairness has gained increasing traction within the ML community in recent years [47]. Within the healthcare field, the need for fair AI models to address existing healthcare disparities and prevent further magnification of biases is particularly important. A landmark study in 2019 found evidence of substantial racial bias in an algorithm used for millions of patients for helping to predict patients with complex health needs [43]. Hence, in order to better understand how well this model performs for individuals with different identities, the fairness of the GNN models is evaluated for two salient characteristics: biological sex and ethnicity. Algorithmic fairness is a nebulous term that encapsulates many different definitions. Three of the most commonly used definitions of fairness are: demographic parity, predictive parity, and equalized odds. To complicate the situation, the impossibility theorem of fairness [26] states that no two of these three metrics can be true at the same time for a well-calibrated classifier and an attribute that could introduce bias.

In our specific task, demographic parity refers to parity in the prevalence of positive mental health predictions for the groups of interest. Since demographic parity only assessees the rate of positive predictions and not the quality of these predictions, this metric will be discarded in favor of the other two fairness metrics. Equalized odds parity is satisfied when the True Positive Rate (TPR) and True Negative Rate (TNR) are equivalent for the groups of interest. TPR is oftentimes referred to as sensitivity, while the TNR is referred to as specificity. Predictive parity, in contrast, is satisfied when the Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are equivalent for the groups of interest. Equations for these four values are shown for reference in Equations 5.3-5.6. For all four metrics, a higher value indicates better performance than a lower value.

$$TPR = \frac{TP}{TP+FN} \tag{5.3}$$

$$TNR = \frac{TN}{TN+FP} \tag{5.4}$$

$$PPV = \frac{TP}{TP+FP} \tag{5.5}$$

$$NPV = \frac{TN}{TN+FN} \tag{5.6}$$

All assessments of equity are conducted on the test dataset using the best performing models from the previous section. The statistical values relating to these fairness metrics are shown in Table 5.3 for equalized odds and Table 5.4 for predictive parity.

Table 5.3 Equalized Odds Model Comparison (Biological Sex)

Model	**TPR Women**	**TPR Men**	**TNR Women**	**TNR Men**
Baseline	.688	.707	.760	.754
SAGE	.690	.665	.780	.782
GAT	.678	.684	.769	.752
MPNN	.740	.741	.733	.718

Table 5.4 Predictive Parity Model Comparison (Biological Sex)

Model	**PPV Women**	**PPV Men**	**NPV Women**	**NPV Men**
Baseline	.343	.315	.931	.941
SAGE	.364	.328	.933	.936
GAT	.348	.307	.929	.937
MPNN	.335	.296	.939	.945

The models assessed in Table 5.3 operate at relatively similar thresholds, effectively balancing the TPR and TNR. There are no easily identifiable performance biases related to biological sex for either predictive parity or equalized odds. All four models achieve relatively equivalent TPR and TNR for men and women. The MPNN stands out as showing exceptional gender parity with nearly equivalent TPR for women and men and an increased TNR for women compared to men. Since the models assessed in Table 5.4 attempt to balance TPR and TNR and the baseline prevalence of mental health diagnoses within the cohort of 14.36%, the corresponding PPV and NPV are quite imbalanced. In practice, this corresponds to a much higher rate of false positives which drive down the PPV. In relation to gender equity, the models again do not demonstrate any easily-identifiable biases, though women generally have a lower NPV but higher PPV in the models. It is refreshing to see that the model achieves relative parity with respect to predictions based upon biological sex, especially given that there are more males than females in the cohort. Similar analyses for ethnicity are shown in Table 5.5 and Table 5.6 for equalized odds and Table 5.7 and Table 5.8 for predictive parity.

Table 5.5 Equalized Odds Model Comparison (Ethnicity Pt. 1)

Model	TPR Asian	TPR Black	TPR White	TNR Asian	TNR Black	TNR White
Baseline	.286	.500	.702	.923	.811	.746
SAGE	.429	.333	.680	.966	.919	.767
GAT	.143	.500	.686	.906	.919	.745
MPNN	.429	.500	.746	.880	.797	.713

Table 5.6 Equalized Odds Model Comparison (Ethnicity Pt. 2)

Model	TPR Mixed	TPR Other	TNR Mixed	TNR Other
Baseline	.769	.667	.799	.810
SAGE	.692	.688	.832	.850
GAT	.731	.667	.799	.846
MPNN	.731	.729	.754	.799

In contrast, the results relating to equalized odds for ethnicity show some disparities in model performance. For instance, the TPR of Asian and Black children within the dataset is substantially lower than that of White children, indicating that the mental health crises of Black and Asian children would be identified less often with clinical implementation of this tool. Although the corresponding TNR for Black and Asian children is higher, it is still disconcerting to see a differential in model performance based upon ethnicity. Trends relating to model performance for children of mixed ethnicity and children of an ethnicity other than those listed are not as easily discernible. It is likely that this difference in model performance for Black and Asian children stems from differences in representation within this dataset, as Asian children comprise only 1.83% of the dataset while Black children comprise only 1.16% of the cohort. Bias in data and model performance may occur when there is unequal representation of groups [27]; further, the smaller amount of data for these individuals leads to greater variance in calculations of fairness metrics. One general solution to these problems is to increase the diversity of the dataset [37], but this approach is not viable in this scenario as this dataset contains all individuals within Wales with social services contact. Hence, up-sampling approaches such as SMOTE [7] could be utilized, though these methods could lead to a trade-off between aggregate model performance and model equity.

Table 5.7 Predictive Parity Model Comparison (Ethnicity Pt. 1)

Model	PPV Asian	PPV Black	PPV White	NPV Asian	NPV Black	NPV White
Baseline	.182	.176	.333	.956	.952	.933
SAGE	.429	.250	.345	.966	.944	.930
GAT	.083	.333	.327	.946	.958	.929
MPNN	.176	.167	.319	.963	.952	.940

Table 5.8 Predictive Parity Model Comparison (Ethnicity Pt. 2)

Model	PPV Mixed	PPV Other	NPV Mixed	NPV Other
Baseline	.357	.274	.960	.958
SAGE	.375	.330	.949	.962
GAT	.345	.317	.953	.959
MPNN	.302	.280	.951	.965

The results in Table 5.7 and Table 5.8 indicate that predictive parity biases related to model performance across ethnicities are also present within these models. For instance, the PPV for White children tends to be higher than the PPV for Black and Asian children. Again, this is partially offset by an increased NPV for Asian and Black children, but the results are troubling. Future work could explore the applications of methods to reduce these biases when making predictions for groups with less data. Moreover, although discrepancies in model performance due to biological sex and ethnicity were explored, this work does not comprise a comprehensive analysis of model equity. For instance, one could also look into other salient characteristics that may affect model performance including data missingness, socioeconomic status, disability status, and intersectional identity. The ensuing chapter will expand upon the GNN models trained and evaluated in this section by delving into issues of missing data and model interpretability.

Chapter 6

Interpretability of Graph Attention Networks and Missing Data Methods

6.1 Motivation

Having evaluated model performance in the previous chapter, I now focus in this chapter upon analysis of GNN interpretability and applying Feature Propagation (FP) in hopes of improving model performance. This chapter thus seeks to further refine GNN methods in order to increase their potential clinical utility.

6.2 Feature Propagation Methods

6.2.1 General Feature Propagation

The GNN layer types assessed in this work impose strict assumptions that all nodes contain full feature sets. However, this assumption rarely holds up in practice, especially within the healthcare domain. Missing data is a rampant problem within healthcare data that often reduces the predictive power and generalizability of developed models [57]. In the SAIL dataset, some nodes are missing up to 50% of their features. Although various techniques like mean imputation can be used, these techniques are not ideally suited for GNNs since they do not take advantage of the graph structure. Rossi et al. proposed a general approach for addressing this problem using an iterative method that they named Feature Propagation (FP) [49]. In this method, unknown features are first initialized to arbitrary values, values are then propagated via the normalized adjacency matrix, and known features are then reset to their ground truth value. These operations are run until convergence [49]. Since there is a

large amount of missing data within the SAIL dataset, I will apply the FP method to attempt to improve model performance for individuals with substantial missing data.

6.2.2 Feature Propagation for Missing Data

FP was implemented by adapting code from the GitHub associated with the paper from Rossi et al. [49]. Distinct advantages of FP include its low computational overhead and efficient convergence. Results for this technique applied to the models from Chapter 5 are shown in Table 6.1. Results relating to the AUROC of different models on the test set are visualized in Figure 6.1.

Table 6.1 Feature Propagation for Missing Data GNN Model Performance

Model	Sim Metric	AUROC Val	AUPRC Val	AUROC Test	AUPRC Test
SAGE	**Baseline**	**.798 ± 8.2e-4**	**.428 ± 1.7e-3**	**.810 ± 1.1e-3**	**.474 ± 1.5e-3**
SAGE	Dual-Loss Auto	.781 ± 8.9e-4	.388 ±2.6e-3	.796 ± 9.1e-4	.416 ± 3.7e-3
GAT	Baseline	.761 ± 3.9e-3	.355 ± 1.4e-2	.769 ± 4.0e-3	.399 ± 1.2e-2
GAT	Dual-Loss Auto	.766± 8.2e-4	.362 ± 2.5e-3	.776 ± 1.0e-3	.377 ± 1.8e-3
MPNN	Baseline	.784 ± 1.8e-3	.422 ± 1.2e-3	.802 ± 1.4e-3	.461 ± 2.0e-3
MPNN	Dual-Loss Auto	.758 ± 3.1e-3	.360 ± 2.6e-3	.769 ± 2.4e-3	.379 ± 5.3e-3

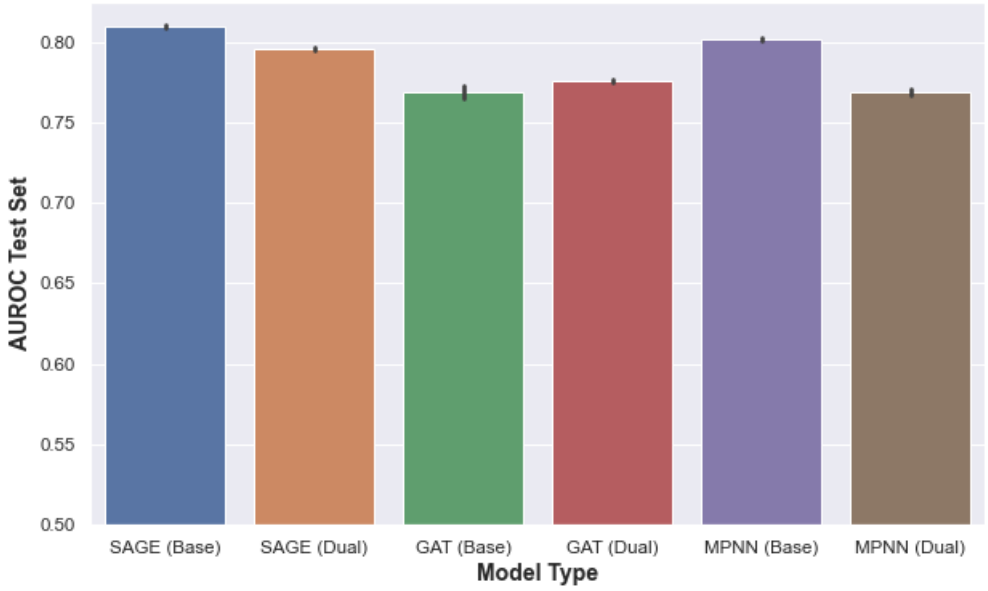


Fig. 6.1 Comparison of AUROC Model Performance on Test Set for Feature Propagation of Missing Data

These models using FP tend to perform comparably or slightly worse than the models without FP, opposing the trend of improved performance found by Rossi et al. [49]. For the SAIL dataset, it is likely that missingness of data is pertinent information that the model should incorporate. Hence, by propagating node feature values from surrounding nodes, this signal regarding the importance of data missingness is smoothed over and lost. This indicates that retaining this information related to data missingness could be conducive to model performance. From a clinical perspective, the importance of missing data lends credence to the idea that patterns of service use could be predictive of mental health status. For example, having numerous healthcare contacts or very few healthcare contacts could both be concerning as an indicator that a child is not doing well. In this vein, neglect could be a potential contributing factor to missingness of data. Moreover, the original datasets upon which FP was evaluated do not incorporate the diverse node features encountered in this dataset. For instance, when propagating missing values for a categorical variable with five different possible values, five different node features corresponding to the one-hot encodings are updated. This creates a new hybrid categorical variable dissimilar from all other non-missing values in the dataset.

6.2.3 Feature Propagation for Output Label Propagation

During the training and evaluation process, the output labels of a node's neighbors are an unused piece of information. Hence, I expanded the use of FP by applying it to output label propagation. In this formulation, an extra node feature is added for each node that I will call the node output feature. To ensure that the model does not have access to the output label when making a prediction for a specific node, the value for that specific node output feature is treated as "missing." Then, the FP algorithm is applied for the node output feature, creating a new representation for that feature which is an aggregation of the node output features of all of the node's neighbors. Since this method requires a graph that changes for each node to predict, it must be implemented using a batch size of one. Results from this method can be seen in Table 6.2. Results relating to the AUROC of different models on the test set are visualized in Figure 6.2. In order to compare model performance across different techniques, results from the best models without FP, with FP for missing data, and with FP for output label are shown in Figure 6.3.

Table 6.2 Feature Propagation for Output Label Propagation GNN Model Performance

Model	**Sim Metric**	**AUROC Val**	**AUPRC Val**	**AUROC Test**	**AUPRC Test**
SAGE	Baseline	.788 ± 2.2e-3	.422 ± 3.8e-3	.803 ± 2.5e-3	.462 ± 2.6e-3
SAGE	Dual-Loss Auto	.763 ± 1.9e-3	.344 ± 3.7e-3	.776 ± 2.3e-3	.367 ± 4.0e-3
GAT	Baseline	.771 ± 6.5e-3	.395 ± 8.8e-3	.785 ± 5.8e-3	.441 ± 7.1e-3
GAT	Dual-Loss Auto	.771 ± 1.9e-3	.371 ± 2.9e-3	.783 ± 1.4e-3	.389 ± 2.1e-3
MPNN	**Baseline**	**.783 ± 2.3e-3**	**.421 ± 5.8e-3**	**.806 ± 2.4e-3**	**.468 ± 5.3e-3**
MPNN	Dual-Loss Auto	.769 ± 3.6e-3	.384 ± 6.0e-3	.782 ± 3.4e-3	.394 ± 9.8e-3

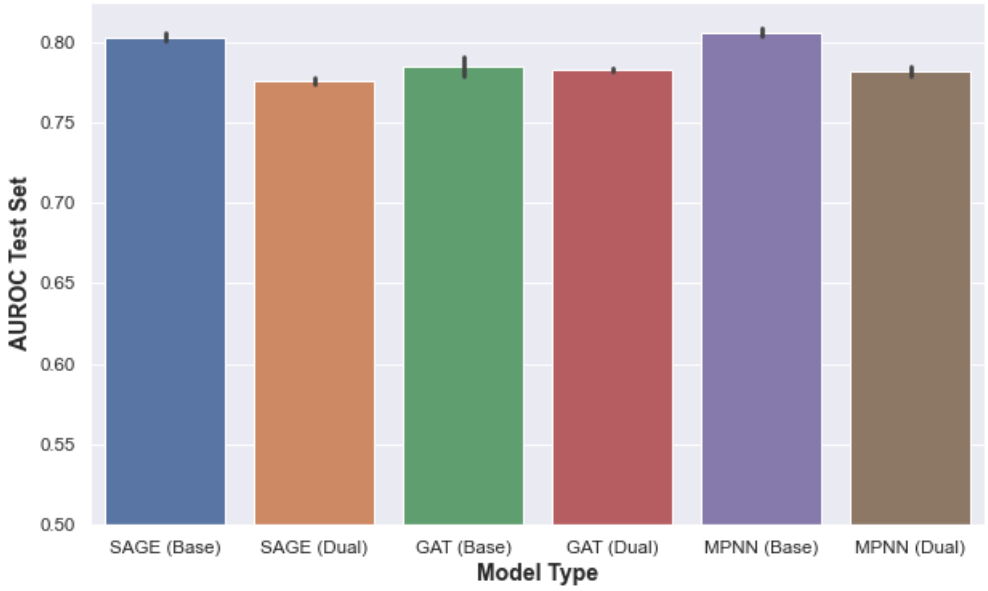


Fig. 6.2 Comparison of AUROC Model Performance on Test Set for Feature Propagation of Output Label

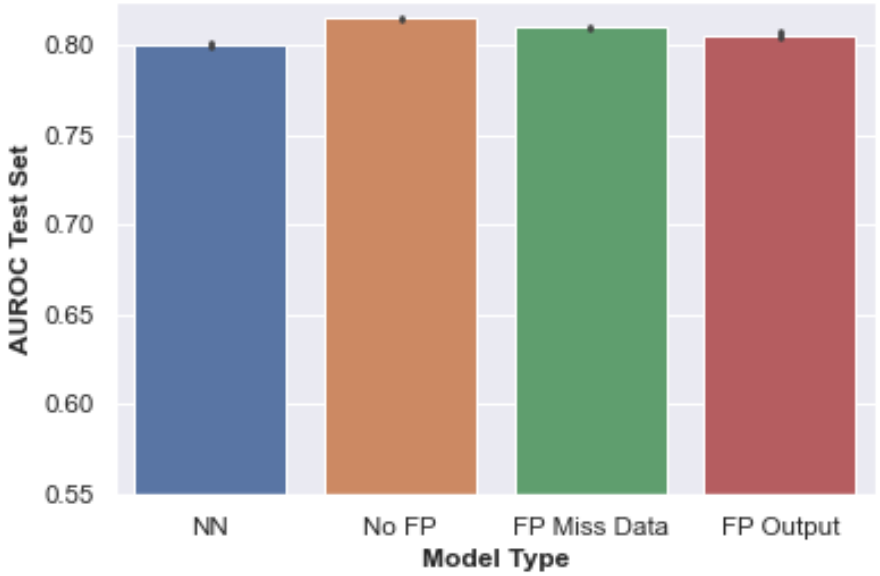


Fig. 6.3 Comparison of AUROC Model Performance on Test Set for Best-Performing Models: Baseline NN (NN), No Feature Propagation (No FP), Feature Propagation for Missing Data (FP Miss Data), Feature Propagation of Output Labels (FP Output)

The best-performing model in Table 6.2 on the test dataset is the MPNN model using the baseline metric for graph construction, which achieves an AUROC of .806. The SAGE model with baseline metric and MPNN model with baseline metric achieved the best performance in all analyses indicating the broad flexibility and solid performance of these two GNN types. In relation to aggregate performance, inclusion of the output label propagation does not affect model performance substantially. Figure 6.3 demonstrates that the best-performing model without FP outperforms both forms of FP. It is important to note that the FP models for output labels did not undergo an extensive hyperparameter tuning specific to the new batch size of one, so it is possible that performance gains could be even greater and the variance in model performance smaller with hyperparameters tuned specifically to this task. Although this FP method for output labels is limiting in that it requires running FP for every node independently, it is worth continuing to explore the additional predictive benefit that this knowledge of the output labels of a node's neighbors provides.

6.3 Graph Interpretability Methods

6.3.1 Motivation

In healthcare settings, model interpretability is a critical prerequisite to clinical adoption. Model interpretability is crucial because physicians feel more comfortable incorporating interpretable models into their clinical decision-making and interpretable models are more amenable to the identification and minimization of pernicious biases and inequities [60]. Like algorithmic fairness, interpretability within ML is a nebulous term generally referring to methods that provide insight into model functioning that help humans understand what is going on inside of the "black box." For the purposes of this thesis, interpretability can be broken down more concretely into two categories: model-level interpretability and instance-level interpretability. Model-level interpretability refers to methods that provide high-level and general explanations for model performance without reference to a specific example. In contrast, instance-level methods provide input-dependent explanations of model outputs by identifying important input features [61].

In this work, I will only address instance-level methods as these methods are most important for clinical decision-making situations in which it is important to understand how a model arrived at a specific decision. Specifically, due to the built-in interpretability of GATs through learned attentional weights [16, 56], I will conduct an analysis of the additional information related to model performance that these attentional weights provide and how the interpretability of these models could be applied to the clinical domain for child mental health

prediction. GATs allow this implicit specification of different weights on neighborhood nodes through self-attention over node features. Thus, for a prediction regarding a patient, GATs provide an interpretable weight for the impact of neighboring nodes on prediction.

6.3.2 Baseline Interpretability

A logistic regression model was trained using the training data to gain insight into model performance and the variables important for model prediction. Model weights for each of the different features are shown in Figure 6.4.

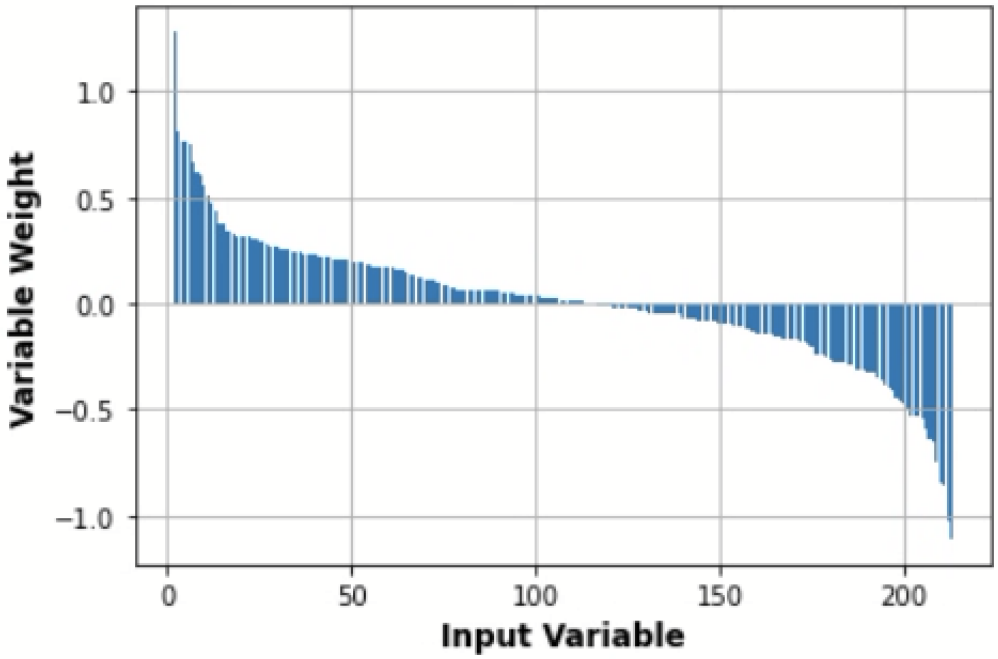


Fig. 6.4 Baseline Variable Importance Analysis (Weights Given by Logistic Regression Model)

No single variable dominates the model, which indicates that a variety of features play a role in the prediction of child mental health. Some of the risk factors that are most predictive of a mental health diagnosis are: substance misuse (largest coefficient), youth offending status, and having a Black or Indian ethnicity. Intriguingly, all of these risk factors relate to characteristics specific to the individual, rather than characteristics that also relate to one's family situation like physical ill health of parents or domestic abuse. A complete exploration of the importance of individual variables is beyond the scope of this study.

6.3.3 Graph Attention Network (GAT) Interpretability

The best-performing GAT model on the test dataset from Chapter 5 was used for all interpretability analyses. I first analyzed how the central node attention weight varies based upon the number of neighbors present in the model. Results from this analysis are shown in Figure 6.5.

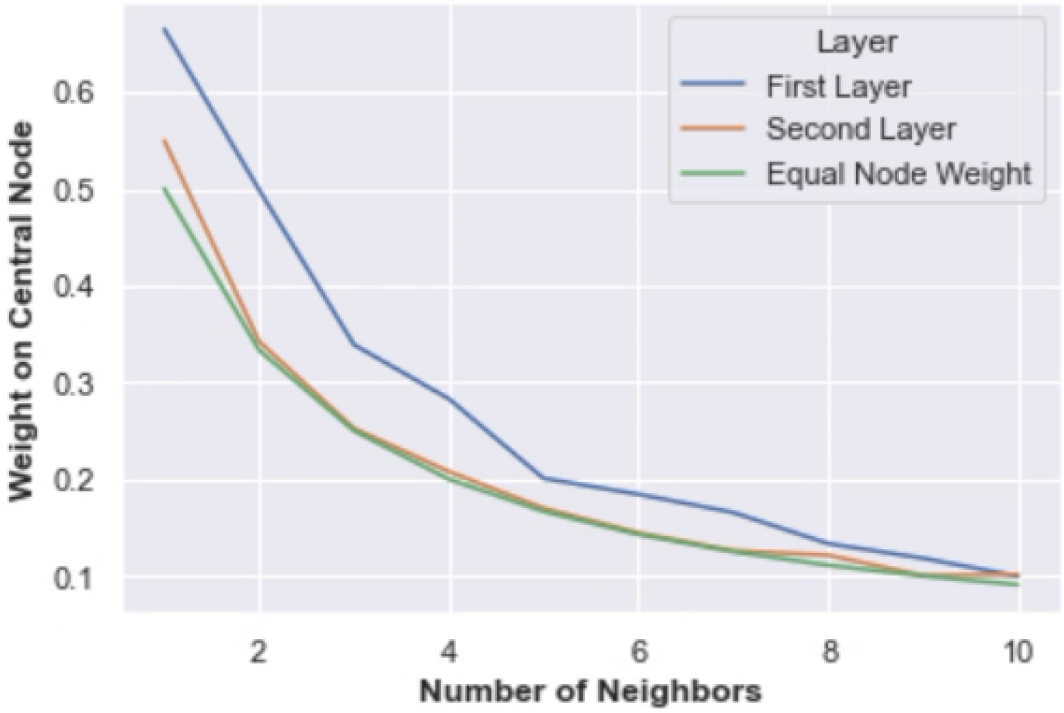
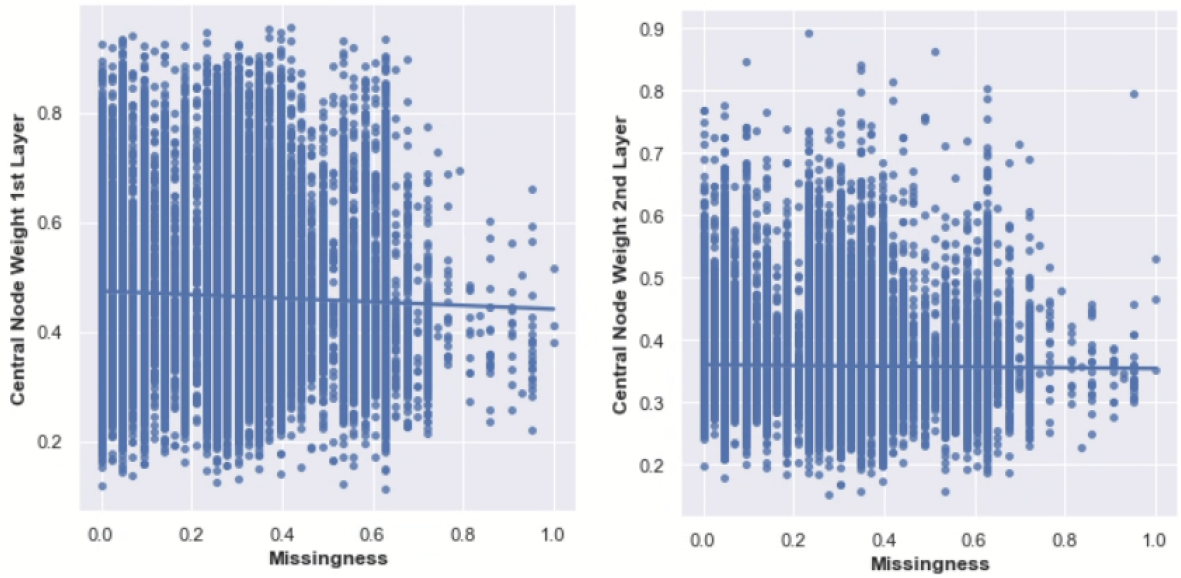


Fig. 6.5 Variability in Central Node Weight of GAT By Neighborhood Size

In Figure 6.5, the blue line refers to the average weight on the central node from the head in the first attention layer, the orange line refers to the average weight on the central node over all heads in the second layer, and the green line shows the trend that one would expect if the weight on the central node was equivalent to the weight of all neighboring nodes. Unsurprisingly, the weight on the central node decreases as more neighbors are added. The second layer generally follows the baseline trend expected with increasing neighbors, while the first layer maintains a higher average weight on the central node. This indicates that the first layer may focus more on the features of the central node, while the second layer with its additional modelling complexity may focus more on features of surrounding nodes.

Additionally, I looked into whether the model is able to deduce a relationship between the quality of data in the central node (represented by the amount of missingness in the node

features) and the corresponding weight on the central node in the first layer. If the model is able to deduce a relationship, one would expect that the model would place less weight on the central node in situations where the node features of that node have higher missingness. An analysis of this relationship is shown in Figure 6.6 with a scatter plot of the variables and a regression model describing the fit of the two variables.



(a) Relationship between 1st Layer Central Node Weight GAT and Data Missingness

(b) Relationship between 2nd Layer Central Node Weight GAT and Data Missingness

Fig. 6.6 Scatter Plot and Regression Line Between Central Node Weight GAT and Data Missingness

In Figure 6.6a, we can see that there is a weak relationship between data missingness and central node weight, with increasing missingness correlating with decreased central node weight. In contrast, in Figure 6.6b, there is not a relationship between the amount of missingness in the node features and the weight of the central node. Hence, the first layer of the model may be discerning a weak relationship between data quality and output certainty, borrowing information from neighbors when the central node data is missing. It is possible that this relationship is not stronger because the missingness of data remains informative for model prediction.

6.3.4 Clinical Utilization of Graph Attention Weights

Graph attention weights provide for the possibility of adding additional interpretable information for individuals making decisions based upon model outputs. For instance, it would

be possible to build a dashboard where physicians could see the result of the model for an individual, the neighbors of that particular individual, and the relative weight that the model afforded to each neighbor. By gaining an understanding of how the model took into account the graph structure of the data to output a label, physicians may gain greater trust in the model. This concept of making decisions regarding patient care based upon previous experiences with similar patients in the past is an integral aspect of how physicians make decisions using their own clinical judgement [24]. This similarity between approaches could lead to greater up-take of these algorithms. It is important to note that GNNs are complicated models, so physicians would require proper educational support regarding ML methods and GNN methods specifically.

Chapter 7

Conclusion and Future Directions

7.1 Synopsis of Salient Points

By exploring the intersection of GNN models and child mental health, I have endeavored to integrate clinical knowledge into the development and evaluation of ML methods. Using similarity metrics derived from patient characteristics in conjunction with GNN models constitutes a broadly applicable technique that can provide a boost in model performance over standard NN methods. Although these methods provide only a modest performance boost for this dataset, it is possible that there could be a greater performance boost in more extensive datasets with lower data missingness and greater time stamp data. Moreover, the results relating to algorithmic fairness based upon biological sex and ethnicity highlight how biases in data quality can affect model performance. Thus, great caution should be exercised and fairness should be evaluated when applying ML methods to datasets where gender or ethnic minorities comprise only a small portion of the dataset population. This caution should be magnified for health datasets where model decisions translate to medical consequences in order to avoid exacerbating existing healthcare disparities. Moreover, my explorations of GAT attentional weights demonstrate that these weights can be helpful for both instance-level interpretability by providing intuitive descriptions of the relative impact of neighbors on model output and model-level interpretability by showcasing how the model functions differently based upon missing data and the number of neighbors.

From a clinical perspective, this work constitutes an important contribution. With the best-performing model achieving an AUROC of .815, these methods merit strong consideration for implementation in a clinical setting. The strong performance of models on this dataset also indicates that integrating educational, social care, and health data is integral to development of a holisitic and effective model. Moreover, despite the importance of early identification of

child mental health problems, there is little work on using ML to predict child mental health with existing studies often limited based upon size or generalizability [39, 19, 53]. In this vein, the linked SAIL dataset is one of the largest and most informationally diverse datasets used for prediction of mental health status in children. Another finding of particular interest to psychiatry is that missing data can itself be predictive. For mental health specifically, it is likely that missing data is heavily tied to patterns of service use and the fact that an individual has fewer contacts with services may be predictive of their mental health status. Hence, in further analyses of datasets related to mental health, great care should be taken to appropriately handle missing data to harness its predictive potential.

7.2 Limitations

It is important to recognize that the generalizability and clinical utility of these findings are affected by a variety of limitations. By explicitly listing out the limitations of this work, I hope that future researchers will be aware of the inherent difficulties and biases associated with the research approach adopted and will be able to identify ways to avoid being encumbered by such limitations. Where applicable, I have also listed the ways in which I have worked to remediate the difficulties associated with these limitations.

7.2.1 Dataset Limitations

The first substantial limitation encountered was the overall quality of the SAIL data. The lack of time stamp granularity hindered our ability to model the data utilizing time-series approaches like RNNs. Hence, I was limited to modelling the dataset utilizing a structure that is static in time, which may have hindered model performance. The quality of the metadata for the various datasets was often quite poor, forcing me to to make assumptions regarding the meaning of values or omit otherwise useful indicators. Additionally, the dataset suffers from the presence of significant missing data. These problems were partially remediated by limiting the cohort to only include individuals with social care encounters and was additionally addressed by exploring various methods such as FP for missing data. It is important to note that, by choosing to focus solely on individuals with social care encounters, the overall generalizability of the model to the entire youth population of Wales is greatly diminished.

In addition to the aforementioned quality of temporal data and data missingness, the SAIL cohort lacks key parental information that could be invaluable for model prediction. Since our dataset contains only children, the models do not have access to maternal data or

other important data on family members. This limited our ability to assess the exposure of children to ACEs as this data is often contained within the health records of parents, rather than the health records of patients. For instance, a study assessing a similar linked database for children found that 72.3% of ACEs were found only in maternal records [54]. Overall, it is important to recognize that the overall performance of these models is inextricably linked to the quality of the data; thus, improving the quality of temporal data and decreasing levels of missingness in future datasets will be integral to overall model performance success.

7.2.2 Modelling Limitations

Specific modelling limitations have been discussed in previous chapters where applicable. Thus, this section seeks to address the broad modelling limitations present within this project. To begin, the project focuses solely upon prediction, rather than clinical utility. Although accurate model prediction is a precursor to clinical utility, high accuracy is not sufficient for a model to be effectively integrated into healthcare practice. Accuracy is one small piece of a more complicated implementation system that depends on factors including model interpretability, model robustness, clinical decision-making, and effectiveness of interventions. Thus, by focusing solely upon model prediction, the utility of this model in a clinical setting is limited without further evaluation.

Another model limitation is that I focused solely on binary classification of presence versus absence of a mental health diagnosis rather than delving more thoroughly into different subtypes of mental health diagnoses. Specific mental health disorders are particularly difficult to diagnose in children given the overlapping symptomology of different disorders and the complications with disentangling general child development from abnormal psychiatric disorders. Although the interventions related to child mental health disorders are somewhat agnostic to mental health subtype, predicting disease subtype could be helpful in increasing the diagnostic potential of this approach. Furthermore, since the outcome labels are likely skewed towards including mental health diagnoses for individuals with more severe mental health problems and groups with greater access to services, it is incredibly difficult to build an equitable classifier. Although I did not extensively explore methods to build more equitable classifiers, I conducted an initial exploration of model equity by assessing model performance for two salient characteristics of individuals (biological sex and ethnicity).

Finally, computational resources were limited. Due to concerns regarding privacy of patient data, SAIL maintains its own proprietary systems for providing additional processing power and RAM. Hence, although money from a grant was allocated towards processing power within SAIL for this project in April 2022, access to the increased processing power

was not provided during the timeframe this study was conducted. This limited my ability to conduct comprehensive experiments addressing all possible combinations of methods.

7.3 Future Directions

The approaches detailed in this work illuminate a variety of intriguing possibilities for future research directions. The most direct path forward for this avenue of research is implementation of similar GNN methods within the CAM-Child database. Since the CAM-Child database will share salient characteristics with the SAIL database, many of the lessons learned from applying ML methods to SAIL will be directly applicable. Although the implementation challenges and details will inevitably vary for the CAM-Child database, it is my hope that this thesis provides key insights into handling linked data for child mental health prediction.

More broadly, this thesis provides a generic framework for building out clinically meaningful patient graphs. The framework involves a process of first utilizing similarity metrics to group together similar patients and then applying a GNN model over the graph structure to make predictions. This framework is adaptable to any number of prediction tasks. Thus, explorations of this generic framework to other datasets where shared characteristics of observations could provide predictive value are merited. In particular, the application of distance metrics to the latent space of autoencoders warrants greater exploration. Further experiments could assess whether variational autoencoders can achieve improved performance. Clustering patients using these similarity metrics is particularly intriguing. Similarity metrics harnessing data relating to risk factors, symptoms, and genetic data could help to identify clinically meaningful and interpretable subgroups of patients that help inform our clinical understanding of complex disease phenotypes. This methodological framework for clustering could be especially powerful for applications within psychiatry given the variability and heterogeneity of psychiatric diagnoses.

Further, this project contains initial investigations into GNN model interpretability with a focus on GAT models. My initial results suggest that attentional weights increase the clinical utility of GATs and more in-depth assessments of the interpretability of Graph Attention Network methods could prove helpful. Specifically, more nuanced analyses of the impact of different input variables on model prediction could improve the interpretability and clinical utility of these approaches. Finally, future work could account for a more nuanced view of mental health diagnoses that looks specifically at subtypes like major depression or splits prediction into other meaningful categories. For instance, one meaningful clinical distinction

that could by explored is delineating between internalizing problems, which are defined by depressive and anxious symptoms as well as somatic symptoms and social withdrawal versus externalizing problems, which are characterized by aggressive or oppositional behavior [58].

7.4 Final Remarks

Reflecting upon this work as a whole, one unifying theme is the interdependent and symbiotic connection between Machine Learning and medicine. In practice, clinical knowledge and ML expertise are often siloed with poor communication between physicians and ML scientists. As we attempt to move towards the implementation of precision medicine and applying ML models effectively and equitably to address systemic biases, this dual clinical and ML perspective will prove invaluable. Through constructive conversations with my two supervisors, I am incredibly thankful to have been privy to these fruitful discussions between ML and clinical experts. In the future, I look forward to continuing to straddle the line between ML and medicine by fostering collaborations with individuals across the two disciplines.

References

[1] Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C., and Petersson, L. (2021). Graph-based deep learning for medical diagnosis and analysis: past, present and future. *Sensors*, 21(14):4758.

[2] Bălan, O., Moldoveanu, A., and Leordeanu, M. (2021). A machine learning approach to automatic phobia therapy with virtual reality. In *Modern Approaches to Augmentation of Brain Function*, pages 607–636. Springer.

[3] Bor, W., Dean, A. J., Najman, J., and Hayatbakhsh, R. (2014). Are child and adolescent mental health problems increasing in the 21st century? a systematic review. *Australian & New Zealand journal of psychiatry*, 48(7):606–616.

[4] Bracher-Smith, M., Crawford, K., and Escott-Price, V. (2021). Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Molecular psychiatry*, 26(1):70–79.

[5] Brady, A. M., Deighton, J., and Stansfeld, S. (2021). Chronic illness in childhood and early adolescence: A longitudinal exploration of co-occurring mental illness. *Development and Psychopathology*, 33(3):885–898.

[6] Butler, A., Van Lieshout, R. J., Lipman, E. L., MacMillan, H. L., Gonzalez, A., Gorter, J. W., Georgiades, K., Speechley, K. N., Boyle, M. H., and Ferro, M. A. (2018). Mental disorder in children with physical conditions: a pilot study. *BMJ open*, 8(1):e019011.

[7] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

[8] Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., and Sun, J. (2017). Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795.

[9] Chung, J. and Teo, J. (2022). Mental health prediction using machine learning: Taxonomy, applications, and challenges. *Applied Computational Intelligence and Soft Computing*, 2022.

[10] Crowley, R. J., Tan, Y. J., and Ioannidis, J. P. (2020). Empirical assessment of bias in machine learning diagnostic test accuracy studies. *Journal of the American Medical Informatics Association*, 27(7):1092–1101.

[11] D'Alfonso, S. (2020). Ai in mental health. *Current Opinion in Psychology*, 36:112–117.

[12] Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

[13] Fletcher, S., Spittal, M. J., Chondros, P., Palmer, V. J., Chatterton, M. L., Densley, K., Potiriadis, M., Harris, M., Bassilios, B., Burgess, P., et al. (2021). Clinical efficacy of a decision support tool (link-me) to guide intensity of mental health care in primary practice: a pragmatic stratified randomised controlled trial. *The Lancet Psychiatry*, 8(3):202–214.

[14] for Mental Health, J. C. P. (2017). Guidance for commissioners of services for people with medically unexplained symptoms.

[15] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR.

[16] Gokden, B. (2019). Coulgat: An experiment on interpretability of graph attention networks. *arXiv preprint arXiv:1912.08409*.

[17] Gupta, A., Liu, T., and Crick, C. (2020). Utilizing time series data embedded in electronic health records to develop continuous mortality risk prediction models using hidden markov models: a sepsis case study. *Statistical methods in medical research*, 29(11):3409–3423.

[18] Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

[19] Haque, U. M., Kabir, E., and Khanam, R. (2021). Detection of child depression using machine learning methods. *PLoS one*, 16(12):e0261131.

[20] Hughes, K., Bellis, M. A., Hardcastle, K. A., Sethi, D., Butchart, A., Mikton, C., Jones, L., and Dunne, M. P. (2017). The effect of multiple adverse childhood experiences on health: a systematic review and meta-analysis. *The Lancet Public Health*, 2(8):e356–e366.

[21] Jensen, P. S., Goldman, E., Offord, D., Costello, E. J., Friedman, R., Huff, B., Crowe, M., Amsel, L., Bennett, K., Bird, H., et al. (2011). Overlooked and underserved:"action signs" for identifying children with unmet mental health needs. *Pediatrics*, 128(5):970–979.

[22] John, A., Friedmann, Y., DelPozo-Banos, M., Frizzati, A., Ford, T., and Thapar, A. (2022). Association of school absence and exclusion with recorded neurodevelopmental disorders, mental disorders, or self-harm: a nationwide, retrospective, electronic cohort study of children and young people in wales, uk. *The Lancet Psychiatry*, 9(1):23–34.

[23] Jones, K. H., Ford, D. V., Thompson, S., and Lyons, R. (2019). A profile of the sail databank on the uk secure research platform. *International journal of population data science*, 4(2).

[24] Kienle, G. S. and Kiene, H. (2011). Clinical judgement and the medical profession. *Journal of evaluation in clinical practice*, 17(4):621–627.

[25] Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

[26] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

[27] Kuhlman, C., Jackson, L., and Chunara, R. (2020). No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv preprint arXiv:2002.11836*.

[28] Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S., et al. (2021). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research*, 23(5):e15708.

[29] Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H.-C., Paulus, M. P., Krystal, J. H., and Jeste, D. V. (2021). Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 6(9):856–864.

[30] Leslie, D., Holmes, L., Hitrova, C., and Ott, E. (2020). Ethics review of machine learning in children's social care. *Leslie, D., Holmes, L., Hitrova, C. & Ott, E.(2020). Ethics review of machine learning in children's social care.[Report] London, UK: What Works for Children's Social Care*.

[31] Lin, E., Lin, C.-H., and Lane, H.-Y. (2020). Precision psychiatry applications with pharmacogenomics: Artificial intelligence and machine learning approaches. *International journal of molecular sciences*, 21(3):969.

[32] Liu, Z., Li, X., Peng, H., He, L., and Philip, S. Y. (2020). Heterogeneous similarity graph neural network on electronic health records. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1196–1205. IEEE.

[33] Liu, Z. and Zhou, J. (2020). Introduction to graph neural networks. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(2):1–127.

[34] Lowthian, E., Anthony, R., Evans, A., Daniel, R., Long, S., Bandyopadhyay, A., John, A., Bellis, M. A., and Paranjothy, S. (2021). Adverse childhood experiences and child mental health: an electronic birth cohort study. *BMC medicine*, 19(1):1–13.

[35] Lucas, G. M., Gratch, J., King, A., and Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

[36] Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., and Gao, J. (2018). Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 743–752.

[37] Madan, S., Henry, T., Dozier, J., Ho, H., Bhandari, N., Sasaki, T., Durand, F., Pfister, H., and Boix, X. (2022). When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence*, 4(2):146–153.

[38] Malone, B., Garcia-Duran, A., and Niepert, M. (2018). Learning representations of missing data for predicting patient outcomes. *arXiv preprint arXiv:1811.04752*.

[39] McGinnis, E. W., Anderau, S. P., Hruschak, J., Gurchiek, R. D., Lopez-Duran, N. L., Fitzgerald, K., Rosenblum, K. L., Muzik, M., and McGinnis, R. S. (2019). Giving voice to vulnerable children: machine learning analysis of speech detects anxiety and depression in early childhood. *IEEE journal of biomedical and health informatics*, 23(6):2294–2301.

[40] Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., and Danese, A. (2022). Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Molecular Psychiatry*, pages 1–9.

[41] Misiak, B., Stańczykiewicz, B., Pawlak, A., Szewczuk-Bogusławska, M., Samochowiec, J., Samochowiec, A., Tyburski, E., and Juster, R.-P. (2022). Adverse childhood experiences and low socioeconomic status with respect to allostatic load in adulthood: A systematic review. *Psychoneuroendocrinology*, 136:105602.

[42] Moriarty, A. S., Meader, N., Snell, K. I., Riley, R. D., Paton, L. W., Chew-Graham, C. A., Gilbody, S., Churchill, R., Phillips, R. S., Ali, S., et al. (2021). Prognostic models for predicting relapse or recurrence of major depressive disorder in adults. *Cochrane Database of Systematic Reviews*, (5).

[43] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

[44] Ochoa, J. G. D. and Mustafa, F. E. (2021). Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patient diagnoses. *medRxiv*.

[45] Ozonoff, S. (2015). Early detection of mental health and neurodevelopmental disorders: The ethical challenges of a field in its infancy.

[46] Pruthi, S. (2022). Mental illness in children: Know the signs.

[47] Richardson, B. and Gilbert, J. E. (2021). A framework for fairness: A systematic review of existing fair ai solutions. *arXiv preprint arXiv:2112.05700*.

[48] Rocheteau, E., Tong, C., Veličković, P., Lane, N., and Liò, P. (2021). Predicting patient outcomes with graph representation learning. *arXiv preprint arXiv:2101.03940*.

[49] Rossi, E., Kenlay, H., Gorinova, M. I., Chamberlain, B. P., Dong, X., and Bronstein, M. (2021). On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features. *arXiv preprint arXiv:2111.12128*.

[50] Ruan, T., Lei, L., Zhou, Y., Zhai, J., Zhang, L., He, P., and Gao, J. (2019). Representation learning for clinical time series prediction tasks in electronic health records. *BMC medical informatics and decision making*, 19(8):1–14.

[51] Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., Steyerberg, E. W., et al. (2021). Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. *Schizophrenia bulletin*, 47(2):284–297.

[52] Solares, J. R. A., Raimondi, F. E. D., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., Gomes, A. C. P., Payberah, A. H., Zottoli, M., Nazarzadeh, M., et al. (2020). Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of biomedical informatics*, 101:103337.

[53] Sumathi, M. and Poorna, B. (2016). Prediction of mental health problems among children using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 7(1).

[54] Syed, S., Gonzalez-Izquierdo, A., Allister, J., Feder, G., Li, L., and Gilbert, R. (2022). Identifying adverse childhood experiences with electronic health records of linked mothers and children in england: a multistage development and validation study. *The Lancet Digital Health*.

[55] Tate, A. E., McCabe, R. C., Larsson, H., Lundström, S., Lichtenstein, P., and Kuja-Halkola, R. (2020). Predicting mental health problems in adolescence using machine learning techniques. *PloS one*, 15(4):e0230389.

[56] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

[57] Wells, B. J., Chagin, K. M., Nowacki, A. S., and Kattan, M. W. (2013). Strategies for handling missing data in electronic health record derived data. *Egems*, 1(3).

[58] Woodman, A. C., Mailick, M. R., and Greenberg, J. S. (2016). Trajectories of internalizing and externalizing symptoms among adults with autism spectrum disorders. *Development and Psychopathology*, 28(2):565–581.

[59] Yang, P.-J. and Fu, W.-T. (2016). Mindbot: a social-based medical virtual assistant. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 319–319. IEEE.

[60] Yoon, C. H., Torrance, R., and Scheinerman, N. (2021). Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics*.

[61] Yuan, H., Yu, H., Gui, S., and Ji, S. (2020). Explainability in graph neural networks: A taxonomic survey. *arXiv preprint arXiv:2012.15445*.

[62] Zhu, W. and Razavian, N. (2021). Variationally regularized graph-based representation learning for electronic health records. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 1–13.

Appendix A

A.1 Best-Performing GNN Models

Details regarding the hyperparameters of the best-performing GNN models are listed below.

1. For the SAGE model using the baseline similarity metric, the learning rate was .01, weight decay was .0005, number of neighbors was 4, size of hidden layer was 128, and the model was run for 275 iterations.

2. For the SAGE model using the dual-loss autoencoder similarity metric, the learning rate was .01, weight decay was .0005, number of neighbors was 2, size of hidden layer was 16, and the model was run for 50 iterations.

3. For the GAT model using the baseline similarity metric, the learning rate was .005, weight decay was .0005, number of neighbors was 2, size of hidden layer was 16, number of heads was 4, and the model was run for 100 iterations.

4. For the GAT model using the dual-loss autoencoder similarity metric, the learning rate was .0005, weight decay was .0005, number of neighbors was 2, size of hidden layer was 16, number of heads was 4, and the model was run for 125 iterations.

5. For the MPNN model using the baseline similarity metric, the learning rate was .01, weight decay was .00005, number of neighbors was 2, size of hidden layer was 32, number of steps was 2, and the model was run for 50 iterations.

6. For the MPNN model using the dual-loss autoencoder similarity metric, the learning rate was .01, weight decay was .00005, number of neighbors was 2, size of hidden layer was 32, number of steps was 3, and the model was run for 75 iterations.