INTEGRATED MULTI-OMICS APPROACH TO PREDICT DEMENTIA: USING AN EXPLAINABLE VARIATIONAL AUTOENCODER (E-VAE) CLASSIFIER MODEL

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE UNIVERSITY OF MINNESOTA

BY:

Sithara Vivek

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF DOCTOR OF PHILOSOPY

ADVISORS:

Dr. Bharat Thyagarajan, MD, PhD, MPH

Dr. Weihua Guan, PhD, MS

April 2023

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to the many people who have supported me throughout my Ph.D. journey.

First and foremost, I extend my heartfelt thanks to my supervisor, Dr. Bharat Thyagarajan, for providing me with the opportunity to work in his lab and for giving me the flexibility to pursue the research I was interested in despite the uncertainties. I would also like to thank both my advisors, Dr. Guan and Dr. Thyagarajan, for their constant guidance and help in polishing my research questions. I cannot thank them enough for their invaluable support in helping me successfully complete my dissertation project. I would also like to acknowledge my other dissertation committee members, Drs. Chad Myers, Eric Lock and Yuk Sham for their support and guidance right from the dissertation proposal writing and providing helpful feedback to improve my dissertation project.

I owe a debt of gratitude to my parents, M.P. Sasidharan Pillai and Rathnamma Sasidharan, for providing me with the best education possible and instilling in me the importance of learning and the qualities of attention to detail and critical thinking, which have been integral to my success. I am forever grateful for their sacrifices and support.

I have been fortunate to have many mentors in my life, beginning with my father, who instilled in me the belief that there are no limits to education and knowledge and motivated me to perform my best. I am deeply grateful to my mentor and advisor in MPH, Dr. Raman Kutty, who taught me critical thinking, the importance of programming languages in research, and helped me secure the best internship position at the Mayo Clinic. I am also grateful to Dr. Miller and Dr. Cooley at the Masonic Cancer Center, who made me realize the need to pursue a Ph.D. to implement my research ideas. I cannot thank Dr. Thyagarajan enough for being an exceptional mentor, advisor, and supervisor who taught me to approach research projects from a different and better perspective and always encouraged me to continue in academia.

I would also like to thank my collaborators, Drs. Eileen Crimmins, Jessica Faul, and Jung Ki Kim, for their invaluable support in obtaining access to the data that I required to complete my projects in a timely manner. I would like to express my gratitude to Drs. Sarah Munro and Ham Lam at the Minnesota Supercomputer Institute for their invaluable support to utilize MSI resources.

Last but not least, I would like to express my heartfelt gratitude to my family. My husband, Vivek Venugopal, has been a constant source of encouragement, supporting me in every possible way to pursue my goals, all while managing our family. I could not complete this journey successfully without his support and I cannot express my gratitude enough. I would also like to appreciate my daughter, Nila Menon, and my son, Niranjan Menon, for their understanding and unwavering support, which helped me complete my Ph.D. successfully. I hope that I have been able to impart upon them the significance of critical thinking and education.

Once again, I would like to extend my deepest gratitude to all those who have supported me in my Ph.D. journey.

Abstract

Alzheimer's disease (AD) and AD-related dementias (ADRD) are complex multifactorial processes where epigenetic and biochemical changes occur many years before the onset of clinical symptoms. During the last decade, large amounts of high-throughput molecular data including genetic variants, and epigenetic and transcriptomic data from blood and brain tissues have improved our understanding of complex molecular mechanisms associated with pathways of AD/ADRD. The application of deep learning methods to analyze integrated multi-omics data may be a powerful approach to elucidate the biological mechanisms in AD. This dissertation aims to develop a framework to process high-dimensional genomics data and to integrate multi-omics data to classify dementia utilizing an end-to-end deep learning classifier model. We developed an end-to-end deep learning explainable variational autoencoder (E-VAE) classifier model, using genome-wide genetic variants (GWAS SNPs) with an accuracy = 0.71 and sensitivity = 0.73 (Chapter 2), and transcriptome (RNA-Seq) with an accuracy = 0.83 and sensitivity = 0.77 (Chapter 3) and epigenetic (DNA methylation) with an accuracy = 0.79 and sensitivity = 0.88 (Chapter 4) collected from 2700 study participants in the Health and Retirement Study (HRS). We utilized a framework to integrate genetic variants and RNAseq data and developed a multi-omics (GWAS SNPs + RNAseq) explainable variational autoencoder (E-VAE) classifier model to predict dementia (Chapter 5) with an accuracy = 0.73 and sensitivity = 0.73. We evaluated the generalizability of the E-VAE classifier models in an external dataset from Religious Orders Study/Memory and Aging Project (ROSMAP) and the multi-omics E-VAE classifier model achieved an accuracy = 0.67 and sensitivity = 0.77. We found that the integrated multi-omics E-VAE classifier model achieved better generalizability in the external data compared to a penalized logistic regression model (accuracy = 0.73 and sensitivity = 0.33) trained using GWAS SNPs and RNAseq. Utilizing the linear decoder in the E-VAE classifier model, we extracted biological interpretable latent features and translated the top-weighted genes into biological insights. We identified genes known to be involved in the pathogenesis of AD/ADRD and novel genes that were not studied previously in association with AD/ADRD. In summary, this dissertation demonstrates the utility of deep learning methods to analyze complex multi-omics data to classify AD/ADRD. The explainable deep learning model, allowed us to interpret the biological importance of deep representations of multi-omics features by optimizing a classifier model for dementia and generating new hypotheses to advance our understanding of the pathobiology of AD/ADRD.

Table of Contents

List of Figures

List of Tables

Abbreviations

Aβ	Amyloid beta
AD	Alzheimer’s disease
ADGC	Alzheimer's Disease Genetics Consortium
ADRD	Alzheimer’s disease and related dementias
ADNI	AD Neuroimaging Initiative
AE	Autoencoder
AI	Artificial Intelliegence
ANN	Artificail Neural Networks
APOE	Apolipoprotein E
APP	Amyloid precursor protein
AUC	Area Under the Curve
BH	Benjamini-Hochberg
BTN	Butyrophylin
CIND	Cognitive Impairment Not Demented
CNN	Convolutional Neural Network
CSF	Cerebrospinal fluid
DEG	Differentially expressed genes
DGE	Differential gene expression
DL	Deep Learning
DMP	Diffentially methylated probes
DNAm	DNA methylation
E-VAE	Explainable Variational Autoencoder

EWAS	Epigenome Wide Association Study
FDR	False Discovery Rate
FHS	Framingham Heart Study
GEO	Gene Expression Omnibus
GFAP	Glial fibrillary acidic protein
GWAS	Genome wide association study
HRS	Health and Retirement Study
HWE	Hardy-Weinberg Equilibrium
LOAD	Late onset Alzheimer’s Disease
LRN	Logistic Regression
LRN	Local Regulatory Networks
MAE	Multi-view Factorization AutoEncoder
MCI	Mild Cognitive Impairment
ML	Machine Learning
MRP	Mitochondrial Ribosomal Proteins
NB	Negative Binomial
NfL	Neurofilament light chain
NIA	National Institute of Aging
NIH	National Institute of Health
NTA	Network Topology Analysis
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PSEN1	Presenilin
qPCR	quantitative Polymerase Chain Reaction

RNAseq	RNA sequencing
ROC	Receiver Operator Characteristics
ROSMAP	Religious Orders Study and Memory and Aging Project
SNP	Single Nucleotide Polymorphism
TCGA	The Cancer Genome Atlas
VAE	Variational Autoencoder
VBS	Venous Blood Study

Chapter 1: Introduction

Alzheimer's disease (AD) and AD related dementias (ADRD) are complex multifactorial processes where the epigenetic, gene expression and biochemical changes occur many years before the onset of clinical symptoms [1, 2]. Over the past few decades, significant progress has been made in understanding the molecular mechanisms underlying AD/ADRD. While a single genetic risk factor for AD has not been identified, there are several genes that have been implicated in the development of the disease, including the Apolipoprotein E (APOE), amyloid precursor protein (APP), presenilin 1 (PSEN1), and presenilin 2 (PSEN2) genes. During the last decade, large amounts of high-throughput molecular data including genomic sequence, DNA methylation and RNA-Seq from blood and brain tissues have improved our understanding of complex molecular mechanisms associated with pathways of AD/ADRD [3]. However, early detection of AD/ADRD remains challenging due to heterogeneity in AD pathogenesis. An integrated multi-omics approach may provide more information about the underlying disease mechanisms [4, 5]. Machine learning methods applied to multi-omics data may be a powerful approach to elucidate the biological mechanisms in AD/ADRD. Deep learning methods have already been successfully applied to various types of cancers and brain diseases for several precision medicine applications and survival predictions. Recent advances in deep learning methods have enabled the identification of new biological insights using heterogeneous multi-omics data with high accuracy for predictions [6, 7].

The overall ***goal*** of this project was to identify a minimally invasive, blood-based multi-component biomarker signature for early detection of AD/ADRD and to understand the molecular mechanisms contributing to AD/ADRD. The primary ***hypothesis*** of this project was that integrated multi-omics-based deep learning classifier model can classify dementia more accurately and robustly as compared to single-omics based classifiers and the deep representations of multi-omics data can be leveraged to identify a robust biomarker profile for early detection of dementia. To achieve this goal, we developed an integrated multi-omics analysis framework by utilizing an end-to-end deep generative neural network prediction model, explainable variational autoencoder (E-VAE) classifier model.

1.1. Background and Motivation

1.1.1. Alzheimer’s disease (AD)/ Alzheimer’s disease related dementias (ADRD)

AD/ADRD are progressive neurodegenerative disorders characterized by memory loss and cognitive decline affecting an individual’s activities of daily living [8]. Currently, more than 6 million US older adults are living with AD [9], and with increases in life expectancy, this number is projected to grow to 14 million by the year 2060 [10, 11]. Studies have shown that dysfunctional interactions of genes and biochemical changes occur many years before the onset of clinical symptoms of AD [1, 2, 12]. The recent transition from symptom-based diagnosis to molecular and pathophysiology-based diagnosis has enabled significant advancement in AD diagnosis, especially with the development of A/T/N based classification scheme using cerebrospinal fluid (CSF) and brain image-based biomarkers [13] focusing on the temporal evolution of biomarkers in AD pathogenesis. Recent evidence from several case-control studies suggests that plasma levels of amyloid, Aβ42/40 ratio and phosphorylated tau proteins [14] correlate with the concentrations of these analytes in the CSF and with imaging biomarkers such as amyloid-PET and tau-PET [15]. In addition, blood biomarkers of neurodegeneration such as neurofilament light chain (NfL) and glial fibrillary acidic protein (GFAP) have also found to be associated with AD progression [16-18]. These plasma protein biomarkers are a new evolving frontier in AD and provide proof that blood-based biomarkers may be helpful in AD/ADRD diagnosis. Still, early identification of AD/ADRD remains challenging due to the progressive and heterogeneous nature of the disease [19]. Integrated analysis of multi-omics data may shed light on cross-talk pattern between functional omics and is necessary to understand the hidden biology in AD pathophysiology.

1.1.2. Multi-omics studies and genetic basis in AD

AD/ADRD has complex etiology with nonlinear and redundant interactions between multiple genes and environmental factors. The pathogenesis of AD/ADRD is not fully understood, but it is known to involve alterations in multiple cellular pathways, including those involved in protein processing, inflammation, oxidative stress, synaptic function, and neuroplasticity. The alterations in these pathways can lead to changes in the genome, epigenome, and transcriptome, which can affect the expression of genes and contribute to the development and progression of AD/ADRD. For example, changes in DNA methylation, histone modifications, and microRNA expression can alter the epigenetic landscape of cells and affect gene expression. Alterations in the transcriptome, such as changes in the levels of mRNA and non-coding RNAs, can also impact the function of cells and contribute to the development of AD/ADRD. Given the complex nature of AD/ADRD, a better understanding of the molecular mechanisms underlying the disease is essential for developing effective treatments and preventative measures. During the last decade, community-based aging studies and consortia focusing on AD/ADRD have generated large amounts of molecular data including genomic variants, epigenome, transcriptome, proteome and metabolome to improve the understanding of complex molecular mechanisms associated with AD/ADRD [3].

The National Institutes of Health (NIH) has launched several initiatives aimed at advancing research into Alzheimer's disease (AD) by collating genetic data from various studies. The Trans-Omics for Precision Medicine (TOPMed) program [20], for example, aims to integrate data from multiple sources, including genomics, transcriptomics, and epigenomics, to better understand the genetic basis of complex diseases like AD. Similarly, the Alzheimer's Disease Genetics Consortium (ADGC) is a collaborative effort that aims to identify genetic variants associated with AD and other neurodegenerative diseases. The GWAS (Genome-Wide Association Study) Catalog is another resource that collates data from studies investigating genetic associations with various diseases, including AD. This resource allows researchers to easily access and analyze genetic data from multiple studies, helping to accelerate research into the genetics of AD. The

AD knowledge Portal is a database repository that collates data from several studies on aging, dementia and Alzheimer's disease that have multi-omics data. It hosts many large collaborative consortium projects aiming to collect and analyze high-dimensional molecular data to study the targets and biomarkers of AD [21]. One specific study in the AD knowledge portal that is relevant to this dissertation is the Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP), ROSMAP (n=3400), initiated in 1994 and 1997 respectively that prospectively collected risk factors of cognitive impairment and measured incidence of AD among nuns/ brothers, priests and laypersons along with a structured neuropathological examination after death [22]. The ROSMAP studies have extensive molecular data that include whole-genome sequencing, DNA methylation, RNA sequencing and miRNA profile along with proteomics and metabolomics data from blood and brain tissues [23]. Large population studies outside of these established consortia are also generating extensive multi-omics data. The Health and Retirement Study (HRS) is an example of a long standing large nationally representative cohort of older adults in the US that has recently generated large amounts of multi-omics data. Since its launch in 1992, HRS has interviewed participants biannually to collect measures of cognitive functioning to determine cognitive decline and onset of cognitive impairment. In 2016, HRS collected a wide range of blood biomarkers as part of the Venous Blood Study (VBS) and performed various molecular assays on biospecimens collected from a subset of VBS study participants. Together, these initiatives and resources are helping to advance our understanding of the genetic basis of AD and accelerating the search for new treatments and interventions for this devastating disease.

Biological systems modeled using high throughput next-generation sequencing data from various facets and biological interactions are necessary to understand complex molecular mechanisms in AD/ADRD. The genetic basis of Alzheimer's disease (AD) includes both rare and common genetic variants. The rare monogenic causes of AD are usually Mendelian, such as mutations in genes encoding amyloid precursor protein (APP), presenilin 1 (PSEN1), and presenilin 2 (PSEN2) [24, 25]. The most well-studied and understood genetic risk factor for AD is the apolipoprotein E (ApoE) gene [26]. In addition to the ApoE gene, genome-wide association studies (GWAS) have identified several other common genetic variants associated with increased risk of developing AD [26, 27], affecting pathways of immunity, lipid metabolism, tau binding proteins, glucose metabolism, mitochondrial function [28] and amyloid precursor protein (APP) metabolism [29, 30]. However, a genome-wide complex trait analysis showed that all known genetic variants can collectively explain only 15% of phenotypic variance in AD [31]. Previous epigenome-wide association studies (EWAS) established the role of epigenetic mechanisms in AD phenotypic diversity ranging from 8.7% from DNA methylation levels of CpG sites in the APOE genomic region in blood [32] to 28.7% from DNA methylation levels of CpG sites in the autopsied brain [33]. An integrated omics network analysis showed that 38 of 364 cerebrospinal fluid (CSF) metabolites can explain more than 60% of variance observed in CSF tau protein [34]. Though studies in brain tissue and CSF are very useful for identifying novel biological determinants of AD, they remain of limited value in population-based screening efforts for the early identification of AD. For such efforts, biomarkers obtained from blood that can be easily obtained from potential patients are essential. However, there is very limited data on the utility of blood-based biomarkers in the early diagnosis of AD. The recent availability of high throughput multi-omics data from blood samples in large population studies such as HRS [35] and other consortiums have brought into the need to develop advanced machine learning methods that will allow us to evaluate the role of multi-omics in AD phenotypic diversity and robustly evaluate the performance of these methods in large cohort studies.

1.1.3. Major challenges in high throughput omics data analysis

The genome-wide omics data has large number of features and has higher order interactions present within and between omics. Due to the high dimensionality of omics data parametric statistical models require huge sample size which is expensive and not always feasible. Biological phenomenon are often stochastic and non-linear in nature driven by combinatorial effect of DNA, RNA, Proteins, Metabolites, Lipids, etc.

resulting into phenotypic variations. A main challenge in multi-omics data analysis is how to efficiently combine information from multi omics data. In the case of Alzheimer's disease (AD), the underlying biology is particularly complex, involving a range of genetic, environmental, and lifestyle factors that can contribute to the development and progression of the disease. The pathological changes in the brain associated with AD, such as the accumulation of beta-amyloid plaques and neurofibrillary tangles, are the result of a complex interplay between various molecules and pathways, including inflammation, oxidative stress, and synaptic dysfunction. While genomics, epigenomics, and transcriptomics provide valuable information about the genome, its modifications, and its gene expression, each regulatory level only contributes to explaining a portion of the variation in phenotype. Multi-omics approaches, on the other hand, can provide a more comprehensive view of the molecular mechanisms underlying complex biological phenomena by integrating information across different regulatory levels. By doing so, they can potentially capture a larger proportion of the variation in phenotype, leading to improved classification and better understanding of biological systems. Another challenge in omics data analysis is the lack of data harmonization and standardization of assays across population-based studies which are essential to develop generalizable ML-based solutions. Data sparsity and missing data are other challenges in genomics data that needs to be taken care of before implementing prediction models.

In order to achieve the objective of personalized medicine for early identification of AD/ADRD, it is necessary to develop innovative analysis methods that can integrate high-throughput omics data from multiple levels. These methods would enable the creation of a minimally invasive approach that could objectively classify individuals with or without dementia. Traditional statistical methods can be limited in high dimensional data analysis especially when there is no prior knowledge to guide the analysis. Also conventional ML models such as penalized logistic regression models and canonical correlation analysis often suffer over fitting in high dimensional omics data due to the complexity and large number of variables involved. These models may be limited in their ability to capture nonlinear relationships between variables.

This brought into the need to develop advanced machine learning methods that can gain new insights into the complex interactions that drive biological phenomena and ultimately develop new strategies for preventing or treating AD. The integrated multi-omics analysis framework we developed account for these challenges and allow us to evaluate the role of multi-omics in dementia phenotypic diversity and robustly evaluate the performance of these methods in large cohort studies.

Applications of deep learning-based models in the classification of dementia phenotypes are still in the primordial stage. A review summarizes the approaches and applications of various integrative methods for multi-omics analysis and demonstrated that single omics analyses using traditional models are not sufficient to characterize a trait [36]. The **holistic approach of integrating hierarchical multi-omics data** may shed light on cross-talk patterns between various functional omics and allow us to understand the heterogeneity of complex data [37]. Studies show that an integrated multi-omics approach provides additional insights regarding the flow of information underlying a disease mechanism compared to a single omics analysis [4, 36]. An integrative analysis of multi-omics data from CSF showed that the major determinants of variation in AD were CSF based AD biomarkers (38.5%), proteins (39.8%), lipids (10.3%), neuro-inflammation markers (10.3%), one-carbon metabolites (9%) and other metabolites (3.7%) [38]. Multi-omics analysis of blood biomarkers in INSIGHT-preAD study demonstrated higher discriminant power for the signature of brain amyloid deposit (AUC=0.994) compared to single omics analysis (AUC =0.881-0.978) among a small number of participants (N=78) [39]. These studies demonstrate that an integrated approach using multi-omics can provide a meaningful classification of individuals with AD/ADRD and those without AD/ADRD [36] and enable precise clinical decision-making [40]. Given, the limited sample sizes of previous studies, an evaluation of combined effect of **genome-wide multi-omics features** [41] is needed for a holistic understanding and robust classification of dementia in older adults. Previous studies have not comprehensively evaluated the role of blood-based multi-omics data that include epigenetic, transcriptomic and genomic variants to classify dementia in a large representative population of older adults. Hence, we propose to develop an integrated multi-omics prediction model using data collected from three omics studies in a large nationally representative cohort of older adults.

1.1.4. Machine learning applications in multi-omics data integration and analysis

Machine learning is a branch of artificial intelligence (AI) mainly focusing on building algorithms to learn to find patterns and features in massive amounts of data to improve predictions and decision making on unseen data. Advancement in machine learning methods using deep learning (DL) architecture allows us to build computational models to extract the information from the complex noisy data [42] and identify the omics match better than the traditional linear methods [6, 7]. DL-based computational models have already found utility in various biological realms to identify disease pathways and modes of action in drug development [43]. Deep learning (DL) methods utilize multiple processing layers to learn representations of data with multiple levels of abstraction [44]. DL methods allow us to **identify intricate structures in large datasets** using the backpropagation algorithm to compute the representation in each layer from previous layers [44]. DL algorithms can handle **batch effects and non-linearity** observed in high dimensional heterogeneous high throughput data while building a computational model for omics data analysis. Moreover, deep learning models can learn from large datasets and can generalize well to new data, making them suitable for analyzing complex biological systems that involve multiple variables and interactions. These models can also identify novel biomarkers and pathways that may not be detected by traditional parametric approaches such as penalized regression models or canonical correlation analysis.

1.1.5. Variational Autoencoders (VAE)

Deep learning is a subfield of machine learning with algorithms motivated by the function and structure of neurons in the brain called an artificial neural network [45]. An artificial neural network (ANN), for ex: an Autoencoder algorithm, consists of multiple layers of calculations featuring an input layer, a hidden layer and an output layer [46]. The deep neural network has multiple hidden layers with each layer successively refines the results of the previous layer in unsupervised learning. VAEs are generative models, a directed probabilistic graphical model with a basic architecture of an autoencoder [47]. Compared to other auto-encoder models, VAE is deeply rooted in a variational Bayesian and graphical model [48]. In a VAE, an encoder is a variational inference net where it maps input data to a probability distribution in the hidden layer, and a decoder is a generative network that maps latent variables back to the reconstructed input data [47]. In VAE, input features are mapped into a distribution instead of a fixed vector and latent variables are the parameterized variables generated from the model which shows specific characteristics of input data [47, 49]. The VAE objective function includes a regularization term that encourages the learned distributions to be close to the prior distribution, usually a standard Gaussian distribution. This regularization term serves to regularize the latent space and prevent overfitting. The smooth and continuous latent space learned by VAEs enables the generation of new data points by sampling from the learned distributions. Compared to commonly used principal component analysis (PCA), VAE uses a nonlinear technique to project high dimensional data to a low dimensional space [50]. Previous studies have shown the utility of VAE models in unsupervised and semi-supervised learning for noise filtering, clustering and anomaly detection and for supervised classification of a trait using latent features [51]. Previous applications of AEs and VAEs in the multi-omics analysis of cancer [52-54] and pan-cancer classifications [55] have been promising and have identified new genome-wide hidden patterns compared to traditional dimensionality reduction methods (PCA). However, the utility of VAE and other DL methods to integrate blood-based multi-omics data to classify AD/ADRD among older adults in large cohorts remains an understudied area of investigation. Hence, we propose to develop an end to end DL model for classification of dementia using the basic architecture of VAE and a classifier network (E-VAE classifier model). The biologically relevant features derived from the proposed VAE model can be used for the early identification of AD/ADRD at its prodromal/ preclinical stages.

1.1.6. The E-VAE classifier model

Training a machine learning model using high dimensional omics data like genetic, epigenetic and transcriptomics suffers from "the curse of dimensionality". We propose an end-to-end DL model with the basic structure of VAE (with an encoder, hidden layers and a decoder) to extract low dimensional feature representations and utilize a classifier network to predict dementia [55]. Using VAE we can generate latent variables, which are parameterized variables created from the model that shows specific characteristics of the input features [47, 56]. These latent variables can be used as input features in the classifier network. To develop the explainable variational autoencoder (E-VAE) classifier model, we have adapted the architecture of a previously published OmiVAE model [55] to implement an unsupervised VAE phase to learn to generate the latent embedding of high dimensional omics data and a supervised classifier network to predict dementia. By implementing a linear decoder layer to generate reconstructed input data from latent features, the prediction model in this study is capable of learning a meaningful latent space that captures the biologically relevant features for AD/ADRD by optimizing a classifier network to predict dementia. This approach can be used to evaluate the difference in feature encoding between dementia categories. The biologically relevant features derived from the proposed E-VAE classifier model can be used for the early identification of AD/ADRD at its prodromal/ preclinical stages.

1.1.7. Applications of variational autoencoders

1.1.7.1. Multi-omics studies in the cancer domain

The architecture of VAE enables direct encoding of biological knowledge using network regularization approaches which will allow for model generalizability in multi-omics data analysis [53, 57]. Ma et.al used the representation learning by Multi-view Factorization AutoEncoder (MAE) and showed that simultaneous integration of feature interaction network and patient view similarity network as constraints into the training objective [53] improved the prediction of cancer subtypes. The OmiVAE model employed an unsupervised task-oriented feature extraction and a supervised classification network as another layer of VAE for multi-class classification in TCGA pan-cancer datasets and demonstrated that the complementary information gained from multi-omics data provide useful biological insights for disease classification than single omics data [55]. Tybalt is a deep learning framework to implement a VAE model that was trained on The Cancer Genome Atlas (TCGA) pan-cancer RNA-seq data and used to identify specific patterns in the VAE encoded features [58].

1.1.7.2. Multi-omics studies in AD/ADRD

Omics for AD classification: Previous studies using machine learning prediction models showed that blood gene expression and methylation data can be used to classify AD [59-61]. Using blood-based transcriptomic data, Lee et al. evaluated various feature selection methods on three public datasets US-based AD neuroimaging initiative (ADNI), Europe-based AddNeuroMed1 (ANM1), and ANM2 and successfully validated gene expression-based AD classifiers across the three study datasets [59]. Another deep learning-based AD classification using a 3D convolutional neural network (CNN) model identified informative features for AD classification from tau positron emission tomography (PET) scan images [62].

Omics for identifying novel pathways and gene networks associated with AD: A Bayesian hierarchical model integrated transcriptomic, DNA methylation and functional gene networks to identify AD candidate genes and networks using the ROSMAP study brain samples [63]. Local regulatory networks (LRN) were used to estimate interactions in local regulatory regions of individual genes and estimate the cross omics interactions from multi-omics data. Tasaki et al utilized this method in the ROSMAP study to estimate the impact of genetic variants and epigenetic variations on gene expression and identify specific neuronal genes that predicted AD dementia [64]. These published studies show proof of principle that ML/DL methods can be useful for dementia classification. However, the utility of VAE and other DL methods to integrate blood-based multi-omics data for dementia classification among older adults in large cohorts remains an understudied area of investigation. Hence, we developed a comprehensive ML prediction model for dementia using VAE.

1.2. Research aims and objectives

The end-to-end deep generative neural network prediction model, the multi-omics E-VAE classifier model, has a basic structure of VAE (with an encoder, hidden layers and a decoder) and a classifier network for supervised prediction. The first objective of our study was (Aim 1A) to develop a framework to incorporate single omics data from genetic (genome-wide SNP data), transcriptomic (RNA-Seq) and epigenetic (DNAm) data for the E-VAE classifier model to predict dementia, and (Aim 1B) to evaluate the biological interpretability of latent features generated from the E-VAE classifier model. The second objective of our study was (Aim 2A) to develop an integrated multi-omics analysis framework to integrate genetic (genome-wide SNP data), and transcriptomic (RNA-Seq) data to develop a multi-omics E-VAE classifier model with a classifier network using learned latent representations to predict dementia, and (Aim 2B) to evaluate the biological interpretability of latent features generated from the multi-omics E-VAE classifier model. In addition, (Aim 2C) we evaluated the generalizability of the E-VAE classifier models in an external study. We used blood-based omics data collected from 2714 study participants in Health and Retirement Study (HRS) to develop the explainable prediction models for dementia and validated the models in ROSMAP study. The overall goal of this project was to identify a non-invasive, blood-based multi-component biomarker signature for early detection of AD/ADRD.

In Chapter 2, 3 and 4, we demonstrate the application of E-VAE classifier model using single omics data to classify dementia in the HRS, validated the generalizability of the single omics prediction models in an external dataset from the ROSMAP study and evaluated the biological interpretability of latent features from single omics feature space. In Chapter 5, we demonstrate the application of E-VAE classifier model to integrate multi-omics data to classify dementia, validated the generalizability of the multi-omics classification model in the ROSMAP study and evaluated the biological interpretability of latent features from multi-omics feature space. In Chapter 6, we summarize the results of single-omics and multi-omics

E-VAE classifier models to classify dementia and show the importance of integrated multi-omics E-VAE classifier model for robust and generalizable classification of dementia.

Chapter 2: Explainable variational autoencoder (E-VAE) classifier model using genome-wide SNPs to predict dementia

This chapter is based on a manuscript that was under peer-review at the time of the dissertation writing.

2.1. Introduction

Alzheimer's disease (AD) and AD related dementia (ADRD) are complex multifactorial neurodegenerative disorders [1, 2], characterized by memory loss and cognitive decline affecting a person's activities of daily living [8]. Studies have shown that dysfunctional interactions of genes and biochemical changes occur many years before the onset of clinical symptoms of AD/ADRD [1, 2, 12]. AD/ADRD does not have a clear inheritance pattern and studies have shown that risk of dementia is polygenic, in that it involves multiple genes with minor effects or the interaction between genes [65]. Identification of susceptibility genes for complex and heterogeneous multi-factorial human diseases has been a significant challenge in human genetics [66]. In the last couple of decades, numerous GWAS studies (146 studies in GWAS catalog) along with large scale consortia powered by meta-analyses of GWAS studies have identified thousands of common genetic variants associated with AD/ ADRD [67]. A large study among Swedish twins also showed the heritability of AD among older adults to be 58% [68]. This explosion of information in human genetics and epidemiology have created new challenges for modeling and interpretation. Challenges in utilizing genome-wide approach for disease classification include the failure of traditional statistical models to account for the higher order interaction in the genome-wide approach [66] and the challenge in selecting useful features for disease predictions from a list of candidate features. Previous studies have developed prediction models for AD/ADRD using Polygenic Risk Scores (based on genome-wide significant loci of AD) or by using relevant Single Nucleotide Polymorphisms (SNPs) (n=50-1000) from GWAS data [69-72]. To address these challenges, we developed a framework using a complex deep neural network model that allows for a large list of features as inputs, accounts for the higher order interactions of features using non-linear model fitting, and enables the interpretation of the importance of features in the context of disease biology.

Recently, a study using ADNI genotype data utilized a deep learning-based Convolutional Neural Network (CNN) approach for SNPs selection and developed a classification model for dementia that achieved an area under the curve (AUC) of 0.82 [73]. Using more advanced deep autoencoder or variational autoencoder algorithm allow us to learn compressed feature representations from high dimensional genome data. In this study, we utilized a deep generative neural network prediction model with a basic structure of Variational Autoencoder (VAE) and a classifier network for supervised prediction. VAEs are a particular type of deep generative model that combines a Bayesian paradigm with a neural network model that can approximate complex, non-linear transformations to learn the latent representation of features from input omics data using variational inference [47, 48]. Thus, unlike the traditional autoencoders, VAEs can handle higher order interactions between high dimensional input features and enhance the classification performance. With advancement in newer machine learning (ML) methods, feature selection algorithms have gained a lot of attention. Feature selection in ML is important to develop optimized algorithms to improve target prediction [74].

In this chapter, we demonstrate the development of a novel end to end deep learning based and explainable variational autoencoder (E-VAE) classifier model, that includes an unsupervised phase using a VAE that generate a latent embedding of the GWAS SNPs and a classifier network that use the learned latent representations to classify a binary outcome. We implemented the explainable model by utilizing a linear decoder layer to extract the learned weights for input features from the trained model. To develop and validate our E-VAE classifier model, we used genotype data from the Health and Retirement Study (HRS) as input features to predict dementia. We used an association-based feature selection method for dimension reduction and a balanced-class outcome to improve the sensitivity of the model. We demonstrated the generalizability of the E-VAE classifier model in an external dataset from the Religious

Orders Study and Rush Memory and Aging Project (ROSMAP) and interpreted the biological functions represented by the latent features from the E-VAE classifier model. To the best of our knowledge this is the first study utilizing a VAE model to use genetic variants to classify dementia and interpret the biological importance of low dimensional latent variable embedding.

2.2 Methods

2.2.1. Explainable Variational Autoencoder (E-VAE) classifier model and its optimization

The E-VAE classifier model is an end to end deep learning model with an unsupervised phase of VAE and a supervised phase of classifier network. Variational autoencoder is a neural network that consists of an encoder, bottleneck layer and a decoder. Compared to classical autoencoders, VAE has a more regularized encodings distribution to avoid overfitting and the specific properties of latent space enable new sample generation/data reconstruction [47]. We adapted the basic architecture of OmiVAE with a fully connected encoder network and classifier network [55] to develop the E-VAE classifier model to classify participants with dementia and we have implemented a linear decoder to extract the weights associated with input features and latent variable embedding by optimizing the model to classify dementia

We considered the GWAS SNPs dataset ‘D’ with ‘N’ samples and ‘m’ omics molecular features, and assumed that each sample $x^i \in R^m$ is generated using latent vector $z^i \in R^p$, where $p << m$. In this process, each latent variable z^i is generated from prior distribution: $p_\theta(z)$ and each sample x^i is generated from the conditional distribution: $p_\theta(x|z)$; and the posterior probability distribution is $p_\theta(z|x)$. VAE use variational Inference (VI), it is a technique of inference in the graphical model, with x the input variable and z the hidden variable. We had to compute the $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$, but the marginal distribution $p(x) = \int p_\theta(x|z)p_\theta(z)dz$ is complex to compute as it is intractable especially in the high dimensional space. So, we approximate a feasible Gaussian distribution $q_\phi(z|x)$ in the probabilistic encoder using VI. Here θ and ϕ are learnable parameters/ weights of the decoder (generative) and encoder (inference) networks respectively.

The Figure 1 illustrates the basic architecture of E-VAE classifier model.

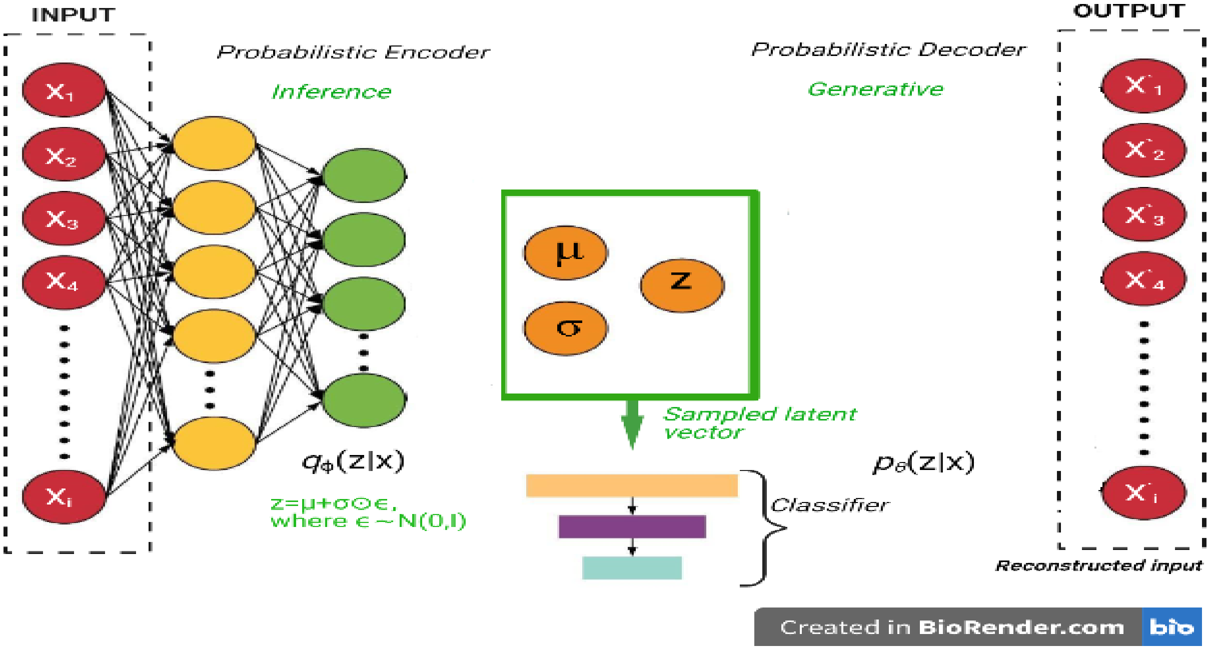


Figure 1

Architecture of the E-VAE classifier model: The diagram showing the basic architecture of the unsupervised phase of VAE with a fully connected encoder layer and a linear decoder and supervised phase using a classifier network to predict dementia. Each node and edges represent a fully connected network. X_i represents each input feature to the inference network and $X`_I$ is the reconstructed input feature from the generative network. The sampled latent vector ‘z’ with the specified latent dimension was then used as input for the classifier network to classify participants in the study.

The loss function of VAE has two terms, the first term to penalize reconstruction error with the objective function to maximize the reconstruction likelihood of input data and the second term to minimize KL divergence regularization with the objective to have the learned distribution $q_\phi(z|x)$ to be similar to the true posterior distribution $p_\theta(z|x)$.

The final **objective function** of VAE is in euquation (1)

$$\text{argmin}\theta,\phi \quad L_{VAE}(\theta,\phi;\, x) = L_{RECONS}(\theta,\phi) + L_{KLD}(\theta,\phi) \qquad (1)$$

$= E_{z\sim q\phi(z|x)}$ [log $p_\theta(x|z)$] - $D_{KL}[q_\phi(z|x)||p_\theta(z|x)]$; where E[,] is mathematical expectation and D_{KL}(,||,) is the KL-divergence term.

In the VAE framework, the encoder and decoder network were jointly optimized by maximizing the evidence lower bound (ELBO) log pθ(x)) ≥ LVAE (θ,ϕ; x). We have minimized the loss in order to maximize the lower bound of the probability of generating real data samples as shown in the equation (2)

$$\log p_\theta(x) \geq E_{z\sim q\phi(z|x)} [\log p_\theta(x|z)] - D_{KL}[q_\phi(z|x) \| p_\theta(z|x)] \quad (2)$$

In the unsupervised phase of E-VAE classifier model, the encoder layer encodes a Gaussian distribution N (μ, σ) with a mean (μ) and a standard deviation (σ) for input sample x_i over the latent space. Then the latent representation z was sampled from the Gaussian distribution $q_\phi(z|x)$. The decoder layer will reconstruct the input sample (x_i`) using latent representations (z) and the VAE loss function will optimize the algorithm.

Reparameterization trick: Sampling from $z\sim q_\phi(z|x)$ is a stochastic process, so we cannot directly backpropagate the gradient. We used the reparameterization trick to use a stochastic gradient to train the VAE network, by multiplying the variance of hidden distribution with a random variable (ϵ) sampled from unit normal distribution N (0,I) and back propagate to identify the parameter of encoder. Then the whole network is differentiable and we can use optimization techniques to solve inference problem successfully. The distribution $q_\phi(z|x)$ is a multivariate Gaussian with a diagonal covariance structure estimated as shown in the equation (3)

$$z\sim q_\phi(z|x) = \mu+\sigma\odot\epsilon, \text{ where } \epsilon\sim N(0,I) \quad (3)$$

The classifier network: The VAE encoder layer learned the latent representations in the unsupervised phase and in the supervised phase the latent variables were used as input to the classifier network to predict dementia. The E-VAE classifier loss was estimated using the binary cross-entropy loss as in equation (4) where ‘y’ was the observed class label (Dementia or Normal) and ‘p’ was the predicted probability of that class label.

$$L_{Dementia} = -1/N \sum_{1}^{N} (y_i . \log(p(y_i) + (1-y_i) . \log(1-p(y_i))) \quad (4)$$

We measured the overall training error using the TOTAL LOSS function of the VAE classification network by combining the VAE (L_{VAE}) loss and the classification loss ($L_{Dementia}$) as in the equation (5) where α and β are the parameters that weight the loss term during training.

$$L_{TOTAL} = \alpha \, L_{VAE} + \beta \, L_{Dementia} \quad (5)$$

The classification network imposed additional regularization for the classification task and learned latent representations by accurately reconstructing the input and to classify the participants with dementia from normal. We used the Adam optimizer to the train the E-VAE classifier model and used Rectified Linear Units (ReLU) activation function and batch normalization to each fully connected block in the encoder hidden layers, a sigmoid activation was used in the linear decoder and a softmax function employed in the final layer of classifier.

2.2.1.1. Hyper parameter tuning: In machine learning, especially when we use deep learning models such as VAE, parameter tuning is a key part of fitting models to data. We tuned the hyper parameters such as epoch, batch size, number of hidden layers, number of nodes in each layer and number of latent features and learning rate using Optuna; a hyperparameter optimization framework. Optuna is a python library that enables us to tune the machine learning model automatically. We optimized the hyperparametrs to maximize the accuracy for prediction using HRS validation dataset along with an early stopping rule for training the model. After hyperparameter tuning, we finalized the encoder network with 2 hidden layers to generate the latent embedding of GWAS-SNPs data based on the model performance on HRS validation dataset.

2.2.1.2. Model evaluation metrics: We evaluated the performance of GWAS-based E-VAE classifier model based on the total accuracy score (fraction of correct predictions) and calculated sensitivity and specificity at the cutoff showing best accuracy. We also used a Receiver-operating characteristic (ROC) analysis to evaluate the non-random prediction performance of the models and used the area under the curve (AUC) to compare the performance of various prediction models [75].

2.2.2. Softwares

We used SAS version 9.4 (SAS Institute, Inc., Cary, NC) and R Statistical Analysis software version 4.0.0 for data preprocessing, data integration and data analysis.

VAE model Implementation: The E-VAE classifier model was built in Python (version 3.8) with PyTorch module (version 1.7.0). We used GPU cluster in Minnesota Supercomputer Institute (MSI) Mesabi system for a faster implementation of the GWAS-based VAE classifier model.

2.2.3. Comparison of VAE model with penalized logistic regression (LR) models

We used penalized logistic regression model, an ElasticNet model, to compare the performance of the E-VAE classifier model. We optimized the penalized LR model by tuning the hyper parameters such as penalty type (l1_ratio) to choose the regularizer and inverse of regularization strength. We used sklearn.linear_model.ElasticNet to implement the penalized LR model.

2.2.4. Biological interpretation of latent embedding

We adapted the method described in Tybalt implementation of VAE [58] model to extract weights from VAE decoder layer to learn the biological signals embedded in the latent features. Due to the complex architecture of encoder layer, we extracted the weights using a linear decoder that captured the contribution of each input SNPs to the learned latent features. We did z-score normalization of the weights and evaluated the distribution of weights of all encoded features to select the positive high and negative high weighted SNPs that had 2.6 SD above or below the mean weight for at least one of the latent feature. We chose the z-score cut-off to have at least 20 -30 genes in the top weighted list. We used geneprofiler2 to convert SNPs to the gene loci. We utilized the WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) to characterize the top weighted genes from the E-VAE classifier model and performed over representation analysis to translate the top weighted gene list into biological insights. We used the Benjamini-Hochberg (BH) method adjusted FDR p-value of 0.05 to select over enriched pathways.

2.2.5. Data

We utilized the genotype data along with cognitive function scores measured in the HRS 2016 survey to build the E-VAE classifier model. The HRS is a biennial survey of nationally representative sample of older adults aged 51 years and older. Since 1992, HRS includes measures of cognitive function to understand the onset and impact of cognitive impairment using an in person and telephone surveys [76]. We also obtained genetic variants data from the Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP), (ROSMAP) from the AD knowledge Portal [21] and accessed the cognitive function measures and basic demographics from RUSH Alzheimer's Data Center to evaluate the generalizability of the E-VAE classifier in an independent dataset. The ROSMAP studies are longitudinal epidemiological and clinical-pathological cohort study initiated in 1994 and 1997 respectively. The ROSMAP study prospectively collected risk factors of cognitive impairment and measured AD dementia incidence among nuns/ brothers, priests and laypersons along with a structured neuropathological examination after death [22].

2.2.5.1 Ascertainment of Dementia in the HRS

Cognitive performance tests in the 2016 HRS survey included measures of episodic memory (using immediate and delayed 10-noun free recall test) and measures of mental processing and working memory (counting backward from 20, serial 7 subtraction). These scores were combined to create a composite score ranging from 0 to 27 that was used to estimate the cognitive function. We used the Langa-Weir (LW) classification approach using the 27 points cognitive function score to group participants into two categories: Dementia (0-6); and Normal (12-27) [76]. Among 2714 participants in the HRS, 128 (4.7%) participants had dementia and 2586 (95.3%) participants had normal cognition in the 2016 survey.

2.2.5.2 Ascertainment of Dementia in the ROSMAP study

A clinical diagnosis of cognitive status was rendered based on a three-stage process including a battery of 19 cognitive tests (comprising 5 domains i.e. episodic memory, visuospatial ability/perceptual orientation, perceptual speed, semantic memory and working memory), a clinical review by a neuropsychologist and a diagnostic classification by a clinician. Clinical diagnosis of AD was based on criteria of the joint working group of the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's disease and Related Disorders Association (NINCDS/ADRDA). Individuals without dementia or mild cognitive impairment (MCI) are categorized as having no cognitive impairment (NCI). Among 234 participants in the ROSMAP study who had both genotype and RNAseq data available, 12 (5.1%) participants had dementia and 222 (94.9%) participants had normal cognition.

2.2.5.3 Harmonization of dementia definition/cognition measures in HRS and ROSMAP study

To optimally harmonize cognition domain scores measured in HRS and ROSMAP using commonly available cognition tests between the two studies, we used z-scores estimated. In the HRS, we used the immediate word recall, delayed word recall, serial 7's test and in the ROSMAP study, we used East Boston immediate recall, East Boston delayed recall and digits backward which were used to measure working memory and episodic memory. We estimated the cumulative cognition score from the three domains and then used the z-score normalization to harmonize the cognition score. We have identified the z-score cut off as '-1.53' in HRS at which there will be no discrepancy with the dementia classification based on published LW classification approach. Similar to the LW classification approach in HRS [76], we classified participants as Normal and Dementia. We used same z-score cut off to classify participants in ROSMAP and compared the classification accuracy using the diagnosis category reported in ROSMAP based on over all cognitive function, neuropsychologist and clinician review. We removed CIND/MCI participants from both studies to keep the outcome class binary. We also removed 7 participants who were classified as having dementia in the ROSMAP study but had HRS z-score based cognitive scores in the normal range.

2.2.5.4 HRS training and validation data with Dementia:Normal class balance

We have included 2714 (Dementia = 128, Normal = 2586) participants from the HRS study. As we had only ~5% prevalence of dementia, we randomly selected 128 participants from the 'Normal' class and the 128 participants with dementia to balance the class distribution to train the E-VAE classifier model. We used a five-fold stratified split to select training and validation data. We combined the first four folds for training data (N=204) and used the fifth fold as validation data (N=52) with equal number of participants in dementia and normal class.

2.2.5.5 HRS GWAS data

The HRS genotyped salivary DNA samples collected since 2006 at multiple time points during field visits yielding a total of 18916 unique participants. We used the most recent version of genotype data from 2006-2012. The genotyping was performed by NIH Center for Inherited Disease Research using the Illumina HumanOmni2.5-4v1/8v1 array and genotyping QC analysis was performed at the University of Michigan using HumanOmni2.5-4v1H for SNP annotation. The Michigan Imputation Server (https://imputationsesrver.sph.umich.edu/) was used for imputation to the 1000 Genomes Project phase 3 integrated variant set (v5, released Oct 2014). The QC filters that were applied to the genotype data prior to imputation included SNP Hardy-Weinberg Equilibrium (HWE) p value $<10^{-4}$, SNP missing call rate ≥2%, and other criteria as specified in Table 1 of the HRS "Quality Control Report for Genotypic Data" [77]. As stated in the chapter by Moore [78], we used a filter approach for feature selection in order to reduce the computational limitation using millions of SNPs. We selected genome-wide significant SNPs (p<5e-08) curated in HRS based on already published large-scale GWAS studies related to ADRD (Gencog:11598 SNPs [79], IGAP_AD:576 SNPs [80], Kunkle_AD_wAPOE:1269 SNPs [80], Wightman_AD:3570 SNPs [81]). We filtered out poor quality and rare SNPs as part of the feature selection. We filtered out SNPs that had low imputation info score < 0.9 and got 14662 unique SNPs from these studies.

2.2.5.6 ROSMAP study GWAS data

To validate the generalizability of the GWAS-based E-VAE classifier model for dementia trained using genetic variants data in the HRS, we utilized the genotype data obtained from blood along with clinical and syndromic phenotype data in the ROSMAP study. In ROSMAP, we had imputed dosage data from two batches of genotype runs; Illumina chop samples and Affymetrix samples with 37 million variants. Based on sample-level quality control assessment, excluded samples with genotype success rate <95% and based on SNP-level quality control assessment excluded SNPs with HWE p-value <0.001, MAF < 0.01, genotype call rate < 0.95, etc. as described in AD Knowledge portal. Population outliers were identified and removed using EIGENSTRAT with default parameters.

2.2.5.7 Harmonization of GWAS SNPs in the HRS and ROSMAP study

We used the unique list of GWAS SNPs in the HRS and matched SNPs from the ROSMAP study genotype data and identified 13,507 SNPs in common. We have filtered out SNPs that were missing in at least one participant in HRS or ROSMAP participants and finalized 5474 SNPs as input features for the E-VAE classifier model.

2.3 Results

2.3.1 Training and validation of the E-VAE classifier model in the HRS

The 204 participants in the training set included 102 dementia cases, 110 women, 113 Whites, with an average age of 73 years. The 52 participants in the validation set included 26 dementia cases, 34 women, 29 Whites, with an average age of 72 years. Table 1 shows the comparison of E-VAE classifier model performance with a penalized logistic regression (LR) model using 5474 SNPs in the HRS validation data and the figure 2 shows the AUC-ROC curves. The E-VAE classifier model achieved an accuracy of 0.71 with 95% CI [0.59, 0.84] and sensitivity of 0.73 in the HRS validation data. We found that the E-VAE classifier had higher predictive accuracy but lower sensitivity in the HRS validation data compared the penalized LR model accuracy of 0.69 and sensitivity of 0.77.

Table 1: Comparison of prediction performances between the GWAS SNPs-based E-VAE classifier model and penalized logistic regression model in the HRS and ROSMAP study.

Models	**Accuracy**	**Sensitivity/ Recall**	**Specificity**	**AUC**
E-VAE classifier model using HRS test (N=52)	0.71 (0.59, 0.84)	0.73 (0.56, 0.90)	0.69 (0.52, 0.87)	0.69
E-VAE classifier model using ROSMAP test (N=234)	0.62 (0.56, 0.68)	0.58 (0.30, 0.86)	0.62 (0.56, 0.69)	0.63
Penalized LR model HRS test (N=52)	0.69 (0.57, 0.82)	0.77 (0.61, 0.93)	0.62 (0.43, 0.80)	0.72
Penalized LR model ROSMAP test (N=234)	0.80 (0.75, 0.85)	0.17 (0.00, 0.38)	0.83 (0.78, 0.88)	0.43

2.3.2 Generalizability of the E-VAE classifier model in the ROSMAP study

The 234 participants in the ROSMAP study included 12 (5%) dementia cases, 170 women, only Whites, with an average age of 84 years. The parameters of the E-VAE classifier model with highest accuracy in HRS validation data was fitted to the ROSMAP data to evaluate the generalizability of the model performance in an independent dataset. The E-VAE classifier model got an accuracy of 0.62 with

95% CI [0.56, 0.68] with an AUC of 0.63 in the ROSMAP data compared to the accuracy of 0.80 with an AUC of 0.43 using the penalized LR model. We found that the E-VAE classifier model had higher generalizability in external dataset compared to traditional LR model. The E-VAE classifier model had better sensitivity to classify dementia in the ROSMAP study, whereas the penalized logistic regression model showed poor generalizability in ROSMAP data.

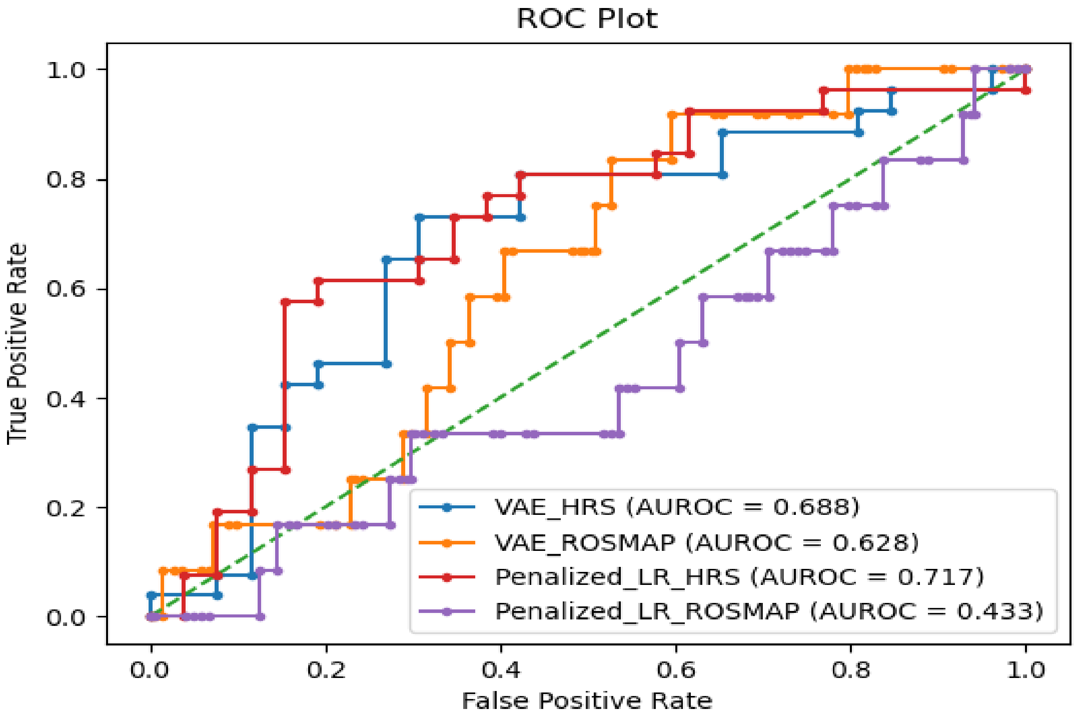


Figure 2: Comparison of AUC-ROC between the E-VAE classifier model and penalized logistic regression model in the HRS and ROSMAP study.

2.3.3 Biological interpretability of the latent embedding of GWAS SNPs

As shown in figure 3, the ORA using Reactome pathways we found that genes in *Butyrophylin (BTN)* family interactions were over enriched in the list of top weighted genes from the E-VAE classifier model. Among top weighted 23 genes, 6 genes were in the *Butyrophylin (BTN)* family interactions gene set of 12 genes in Reactome (https://reactome.org/PathwayBrowser/#/R-HSA-8851680). We did not observe any pathways enriched using KEGG pathways in the ORA analysis.

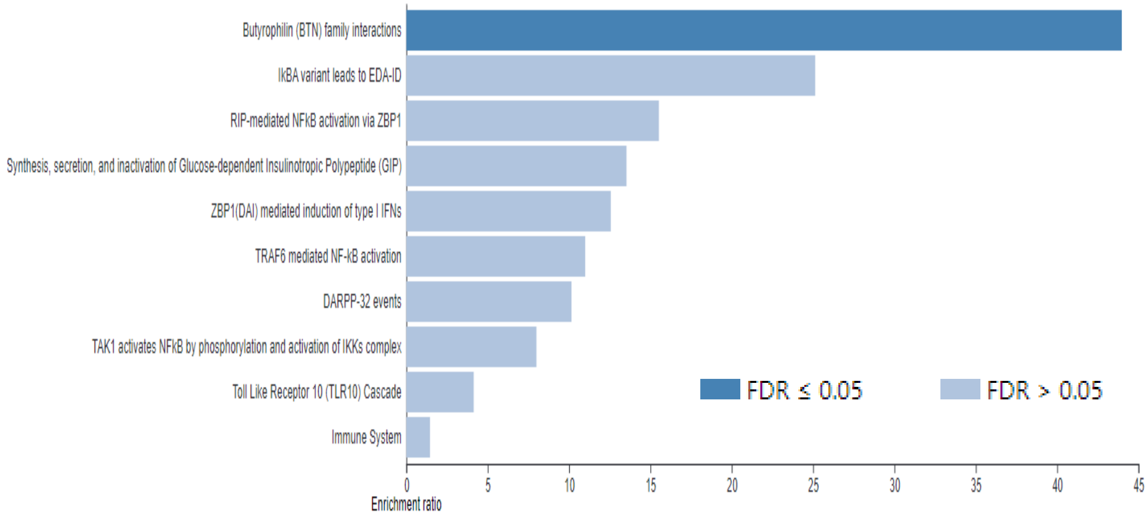


Figure 3: Results of over representation analysis (ORA) using Web based gene set analysis toolkit (WebGestalt).

This figure shows the pathway enrichment results of ORA using Reactome. We used 120 unique entrezgene IDs to annotate to the selected functional categories and for enrichment analysis. We used the Benjamini-Hochberg (BH) method adjusted FDR p-value of 0.05 to select over enriched pathways.

2.4. Discussion

In this study, we used an end-to-end deep learning (DL) model with the basic structure of VAE, a probabilistic Autoencoder (AE) model [47], to extract low dimensional feature representations from genetic variants and utilize a classifier network to predict dementia [55]. The GWAS-based E-VAE classifier model achieved an accuracy of 0.71 with an AUC of 0.69 in the HRS test data and achieved a generalizable accuracy of 0.62 with an AUC of 0.63 in the ROSMAP study data. We found that the E-VAE classifier model had better generalizability in external dataset compared to a penalized logistic regression model. We demonstrated the utility of a linear decoder to extract the input feature weights for the latent embedding of

SNPs for the biological interpretability by optimizing a classifier model for dementia. We found that genes from *Butyrophylin (BTN)* family interactions gene sets were over-represented in the Reactome pathways.

Genome-wide association studies (GWAS) of AD have identified many new genetic loci [26, 27], affecting pathways of immunity, lipid metabolism, tau binding proteins, glucose metabolism, mitochondrial function [28] and amyloid precursor protein (APP) metabolism [29, 30]. However, a genome-wide complex trait analysis showed that all known genetic variants can collectively explain only 15% of the variance in AD even though heritability of AD has been estimated to be 58% in twin studies [31, 68]. Generally, predicting neurodegenerative disorders are challenging as the AD pathological information are not easily accessible [82]. To address this issue, developing computational methods to identify novel gene candidates from saliva or blood, easily accessible biospecimens, is necessary for early detection and targeted intervention of AD/ADRD. A deep learning-based Convolutional neural network (CNN) model developed by Jo et.al using ADNI genotype data showed the power of deep learning approaches using genetic variants to predict AD [73]. They developed a three step approach to identify phenotype associated SNPs and to develop a classification model. They mainly identified the well-known APOE region as most significant locus for AD [73] using the phenotype information score. Compared to their reported accuracy of 75% using a CNN model with 4000 SNPs, we obtained a similar accuracy of 71% with 5474 SNPs in the HRS test data. However, the previous study did not evaluate the validity of their model in external datasets. We have demonstrated that our findings had a 62% accuracy in an external dataset. Previous studies have showed that association-based feature selection from the training data can improve prediction performance of ML models [83] and this may partly account for the high accuracy reported with the CNN-based model. In contrast, our study did not select SNPs based on the association in the HRS dataset alone and instead used a list of SNPs that were associated with AD/ADRD in external datasets independent from the datasets used in this analysis. Another advantage of our study is that the GWAS-based E-VAE classifier model we developed is capable of learning a biologically relevant latent space.

A recent study using genotype data compared the predictive performance of different ML models [69] and showed that LASSO model performed the best with SNPs (n=1106) associated with misclassified samples included in the feature pool and achieved an AUC of 0.84. Major limitations of this study were the inclusion of samples from the test dataset for the feature selection and lack of validation of the model performance in an external study. Compared to their feature set, our dataset included much larger number SNPs that could account for the variations in risk of dementia among older adults. We showed that GWAS-based E-VAE classifier model had better generalizability to external dataset compared to traditional penalized logistic regression models. Recently, You et.al developed a dementia prediction model coined as UKB-DRP model using top 10 predictors from 366 candidate variables including genetic, environmental and physical measures collected in longitudinal population based study using UK biobank data achieved an AUC of 0.85 [71]. They used an inner-looped cross validation for hyper parameter tuning and evaluated the model performance in a validation set. The UKB-DRP model implemented using LightGBM algorithm will work only for studies with large sample size and high feature space like UK biobank database and it lack validation of model performance in an external cohort. They identified Age and APOE e4 as major predictors of dementia. We might be able to improve the classification performance of GWAS-based E-VAE classifier model considering APOE and social determinants of health.

Applications of DL-based models in the classification of AD/ADRD phenotypes are still in the nascent stage. Recent advancements in deep learning methods allow us to analyze high throughput data generated from GWAS studies in AD [84] to identify novel biomarkers for early detection. Though GWAS played an important role in identifying the 'candidate genes' or 'candidate loci' susceptible for risk of AD, this information will not be complete if we cannot characterize the functional context of these candidate genes in relation to the pathways or other knowledge structures that define the cause-effect relationship [82]. To address this, we have adapted the linear decoder method to extract the decoder weights to identify the contribution of the GWAS SNPs to the learned latent features by optimizing a dementia classification network. Among the top weighted genes identified using the GWAS-based E-VAE classifier model, the genes from *Butyrophylin (BTN)* family interactions gene set were over-represented in Reactome pathways. Butyrophilins (BTNs) are a group of the moonlighting proteins in immunoglobulin family. Studies have explored the role of butyrophilin-related molecules to regulate T-cell activation [85] and found to be associated with pathology of many human diseases [86]. The biologically interpretable latent features from the E-VAE classifier model allowed us to explore the complex molecular mechanisms in AD/ADRD etiology and generate new hypothesis to study.

2.4.1 Strengths and limitations

A prediction model trained using omics data from one study and validated in an independent study is hard to implement in large cohort studies due to missing data, the technical variation in the measurement of omics data, and differences in the measurement or definition of the outcome variables. In this study, we leveraged results from four large-scale GWAS studies to inform our SNP selection. Since the HRS and ROSMAP employed identical methods for measurement of genetic variants we have implemented a generalizable and explainable E-VAE classifier model for dementia using the GWAS SNPs. The imbalanced datasets severely limit the generalizability of the model as it bias the model to learn to predict the larger class. To overcome this we utilized a balanced-class approach to train the E-VAE classifier model. A major limitation of our study is the small sample size and inability to account for all the heterogeneity observed in genetic studies of AD. Future studies with larger sample sizes can account for the heterogeneity and increase the predictive accuracy of the GWAS-based E-VAE classifier model. Another limitation is the definition of dementia used in the HRS and ROSMAP study. The HRS surveys were tailored to identify factors contributing to the development and characterization of cognitive impairment and dementia among older adults of 51 years and older. Compared to the HRS, the ROSMAP study included more targeted population of older adults and captured their cognitive trajectory using a detailed battery of neuropsychological tests and neurological evaluation. We used 3 domains of cognition measures common to both studies and classified participants using a harmonized dementia definition. Also we did not include already known risk factors such as age and APOE allele status in the E-VAE classifier model to predict dementia.

2.4.2 Innovation

Though VAE models have been applied successfully in other health domains such as cancer, the application of VAE model using SNPs in the classification of AD/ ADRD is novel. The biological interpretability of latent features generated from a dementia prediction model is an innovative approach. This allowed us to translate the deep learning-based latent features to characterize the genetic etiology of dementia. In addition, the interpretation of the biological meaning of latent features may lead to new biological hypothesis to be investigated. Availability of genetic variants data from multiple studies of large representative population of older adults allowed for better generalizability of study results.

2.4.3 *Conclusion*

Dementia etiology is complex and is influenced by nonlinear and redundant interactions of genes, when a specific pathway alters from the normal these changes are captured in a genome, epigenome and transcriptome. Utilizing a GWAS-based E-VAE classifier model optimized to predict dementia, we have extracted a biologically relevant latent representations from high dimensional genetic variants data. The E-VAE classifier model accounts for the higher order feature interactions and noise in genome-wide data in the high dimensional space compared to classical dimensionality reduction algorithms. The knowledge gained from this study can be further validated to help develop blood-based signatures for the early identification of AD/ADRD in older adults. Additional functional genomics studies are necessary to identify the genetic pathways to AD/ADRD with the integration of transcriptomic and epigenomic data.

Chapter 3: Explainable variational autoencoder (E-VAE) classifier model using RNAseq data to predict dementia

3.1. Introduction

Alzheimer's disease is a progressive neurological disease with a multifactorial complex pathology. The molecular mechanism of AD has not been explored well specifically lacking a genetic risk factor for AD. In the past decade, our comprehension of intricate molecular mechanisms linked to AD pathways has been enhanced due to the availability of significant volumes of high-throughput molecular data such as genetic variants, DNA methylation, and RNA-Seq obtained from blood and brain tissues [3]. Though the cerebrospinal fluid (CSF) and neuroimaging biomarkers were proven to be effective for AD diagnosis, the highly invasive nature and high cost of these biomarkers limits its application in regular clinical examination for early identification of AD. Also, these biomarkers are useful only when the symptoms are evident. Compared to these biomarkers, the blood-based biomarkers are getting more attention due to less invasive and affordable nature for the diagnosis of AD. For such efforts, biomarkers obtained from blood that can be easily obtained from potential patients are essential. However, there is very limited data on the utility of blood-based biomarkers in the early diagnosis of AD. The recent availability of high throughput omics data from blood samples in large population studies such as HRS [35] and NIH initiatives such as Trans-Omics for Precision Medicine (TOPMed) program [20] have brought into the need to develop advanced machine learning methods that will allow us to evaluate the role of multi-omics in AD phenotypic diversity and robustly evaluate the performance of these methods in large cohort studies.

3.1.1. Blood based biomarkers of AD

Pilot case-control studies identified the importance of plasma phosphorylated tau (p-tau181, p-tau231) as promising biomarkers of brain AD pathology [14, 18, 87] and a prospective study showed that lower ratio of Aβ1-42/Aβ1-40, and higher levels of p-tau181 and GFAP in plasma were associated with increased Aβ-PET load [88]. Concentration of amyloid and phosphorylated tau proteins in blood corresponds with the concentrations of those in CSF and with amyloid-PET and tau-PET scans [88] and blood biomarkers of neurodegeneration such as neurofilament light chain and glial fibrillary acidic protein were found to be correlated with disease progression [89]. Given the breakthrough of p-tau families as potential blood biomarker for the early detection of AD, interrogation of the whole-genome wide transcriptomic data to understand the hidden biology in AD/ADRD and uncover mechanisms of action for therapeutic development would be of great value. To support this idea, a study using the Multi-task Deep learning for Alzheimer's Disease neuropathology (MD-AD) framework analyzed both brain and blood gene expression data, and despite variations in gene expression levels, the latent embedding of blood samples was able to accurately predict neuropathological phenotypes corresponding to the cognitive status of the donors [90].

Protein-coding genes are the major part of annotated genes in the human reference genome so we can understand their biological functions. However recent studies in AD emphasized the need to evaluate both protein-coding and non-coding genes [84, 91, 92]. Comprehensive analysis of whole genome-wide blood transcriptome data using microarray and next generation RNA sequencing methods have been widely applied to identify differentially expressed genes in particular phenotype . A meta-analysis of microarray datasets from blood transcriptome data identified several dysregulated genes and pathways in the pathogenesis of MCI and ADRD [93, 94]. Microarray-based gene expression profiling has been established to identify potential blood biomarkers of AD, a small study (N=184) among AD patients in Malaysia identified novel biomarkers of AD using a feature selection method Boruta's algorithm [95] and training an elastic net logistic regression model on transcriptome data [96]. Another small study (N=180) using AD dataset from the Gene Expression Omnibus (GEO) repository showed high accuracy of prediction models using blood transcriptome data and suggested that features of oxidative stress induced by β-amyloid were the key features in their model [97]. A large scale RNAseq blood transcriptome study performed differential gene expression (DGE) analysis and identified hub genes associated with the pathogenesis of AD that could be potential biomarkers of AD [98].

Deep learning methods have been applied successfully to investigate bulk-RNAseq and scRNAseq data. Numerous studies have developed deep learning architectures to classify AD/ neuropathological phenotypes using brain transcriptome data [90, 99]. Among these studies, the MD-AD framework utilized learned representations from brain gene expression data that can span to blood tissue gene expression data and successfully predicted participant's neuropathological phenotypes to match their cognitive status [90]. These studies support utilizing advanced deep learning neural network-based models to successfully analyze high dimensional transcriptomics data. Here we propose to utilize a modified Variational Autoencoder based classifier model as described in Chapter 2 (section 2.2.1). Compared to the conventional ML models, VAE models account for the nonlinear interactions of high dimensional transcriptomics data. Unlike traditional VAE approaches, the E-VAE classifier model we developed enforces additional regularization using a classifier network to predict dementia from the learned latent features.

In this chapter, we demonstrate the application of the framework to incorporate transcriptomics data for the E-VAE classifier model to predict dementia and to evaluate the biological interpretability of latent features generated from model. The VAE model in our proposed study is capable of learning a meaningful latent space that captures the biologically relevant features for AD. This project highlights the importance of identifying biologically interpretable deep transcriptomic features from high dimensional RNAseq data to explore the complex molecular mechanisms in dementia etiology and utilize those for early identification of disease.

3.2. Methods

3.2.1. E-VAE classifier model and its optimization

We have adapted the architecture of E-VAE classifier in Chapter 2 to develop an RNAseq-based E-VAE classifier model to classify dementia. Briefly, the E-VAE classifier had one hidden layer in the encoder network and a linear decoder with the classifier network to predict dementia. We considered the gene expression dataset ‘G’ with ‘N’ samples and ‘g’ omics molecular features, and assumed that each sample xi ∊ G^g is generated using latent vector zi ∊ G^p, where p << g. We used the total loss function to estimate the training error by combining the VAE (L_{VAE}) loss and the classification loss ($L_{Dementia}$) as in the equation (5) in Chapter 2.

$$L_{TOTAL} = \alpha\, L_{VAE} + \beta\, L_{Dementia}$$

We used Optuna to tune the hyper parameters such as epoch, batch size, number of hidden layers, number of nodes in each layer and number of latent features and learning rate to optimize the RNaseq–based E-VAE classifier model. The model evaluation metrics and implementation were same as in Chapter 2. We used a penalized logistic regression model to compare the performance of the RNAseq-based E-VAE classifier model.

3.2.2. Biological interpretation of latent embedding of RNAseq features

As described in detail in Chapter 2, we implemented a linear decoder to extract the input feature weights for each latent feature and interpreted the biological meaning of embedded features by optimizing a classifier model to predict dementia. We z-score normalized the weights and evaluated the distribution of weights of all encoded features to select the positive high and negative high weighted SNPs that were 2.2 SD above or below the mean weight. The cut off was chosen based on a natural selection of 20 – 30 genes for biological interpretation.

3.2.3. Data

The HRS has initiated a sub study Harmonized Cognitive Assessment Protocol (HCAP) which is part of HCAP network, an ongoing international research collaboration funded by the National Institute on Aging (NIA) that seeks to measure and understand dementia risk by collecting a carefully selected set of established cognitive and neuropsychological assessments and informant reports to better characterize cognitive function among older people. The HRS-HCAP include a random sample of approximately 4,000 HRS respondents with venous blood samples collected as part of Venous Blood Study (VBS) and performed innovative assays RNA-seq, and DNA methylation. We have included 2714 participants who had GWAS SNPs, DNA methylation and RNAseq data available in the HRS to develop the RNAseq-based E-VAE classifier. We used the outcome 'Dementia' defined in Chapter 2 to develop the RNAseq-based E-VAE classifier model and employed same data harmonization between the HRS and ROSMAP study as described in section 2.2.5 of Chapter 2.

3.2.4. HRS RNA extraction, cDNA library preparation and RNA sequencing

RNA extraction: Total RNA was extracted from PAXgene tubes with the PAXgene Blood RNA Kit IVD from Qiagen Inc. (San Diego, CA). Extracted RNA is then stored at -80°C until further analysis. Total eukaryotic RNA isolates were quantified using a fluorimetric RiboGreen assay. Total RNA samples were treated with the Globin-Zero Gold rRNA Removal Kit (Illumina Inc.) and were converted to Illumina sequencing libraries using Illumina's stranded mRNA Sample Preparation kit (Cat. # RS-122-2101). One microgram of total RNA is oligo-dT purified using oligo-dT coated magnetic beads, fragmented and then reverse transcribed into cDNA, fragmented, blunt-ended, and ligated to indexed (barcoded) adaptors and amplified using 15 cycles of PCR. Indexed libraries were then normalized, pooled and size selected to 320bp +/- 5% using Caliper's XT instrument.

RNA Seq: All 2714 participant samples were sequenced as 2*50 bp paired-end sequences with a minimum of 20 million reads per sample on NovaSeq. All samples were processed through the HRS RNAseq QC analysis pipeline at the University of Minnesota, this is an extended version of the TopMed/GTEX analysis pipeline (https://github.com/broadinstitute/gtex-pipeline/blob/master/TOPMed_RNAseq_pipeline.md). The STAR aligner was used for alignment of the sequence reads to the GRCh38 human reference genome along with GENCODE 30 annotations. All quality control analyses were performed using an updated version of RNASeQC 2.3.4 and estimated quality control metrics to obtain the final data. The read counts from each sample were combined into a count file.

Normalization/Transformation: We used edgeR [100] calcNormFactors() function and used RLE (relative log expression) normalization to account for compositional differences between the libraries. RLE is the scaling factor method; where the median library is calculated from the geometric mean of all columns and the median ratio of each sample to the median library is taken as the scale factor. We then used cpm() function in edgeR on the normalized DGEList object to estimate the log2 counts-per-million (log2cpm) with a prior.count=2.

Filtering out low expressed genes: We filtered out poorly expressed genes to reduce the burden on multiple testing to estimate FDRs. We included genes with log2cpm values >=3 in at least 10% of the participants. The log2cpm of 3 was used as a cut-off based on the blinded duplicates analysis to reduce the noise level from low expressed genes.

Differential gene expression (DGE) analysis: After estimating negative binomial (NB) dispersion of the data, we fitted gene-wise negative binomial GLMs to determine differentially expressed genes. Then we used likelihood ratio tests for Dementia vs. cognitively normal (control) samples to identify top differentially expressed genes. All p-values were corrected for multiple testing using the Benjamini-Hochberg fasle discovery rate (FDR) method [101]. We used MA plot (an application of a Bland–Altman plot for visual representation of transcriptome data) between the average log2CPM of the genes and the log2fold changes to visualize the distribution of DEGs. We selected the features at FDR cut-off of 50% by multiple iterations of E-VAE classifier model.

3.2.5. ROSMAP RNAseq data

Monocyte isolation and sorting: After thawing, PBMCs were washed in 10ml PBS (Lonza, 17-516F) and CD14+CD16- monocytes were isolated using the EasySep Human Monocyte Isolation Kit (Negative selection kit, Stemcell Technologies, 19359) according to manufacturer's instructions. The monocyte suspension was spun down and incubated with anti-CD14 AlexaFluor488 (Biolegend, 301804, clone M5E2) and anti-CD16 APC (Biolegend, 302012, clone 3G8) antibodies as well as Live Dead Fixable Aqua Dead cell stain (Invitrogen, L34957) for 20 min on ice. Subsequently the cell suspensions were washed twice with staining buffer (PBS containing 1% FBS (Gibco, 10438026)), filtered through a 70 µm filter and the Live (BV510-) CD14+/CD16-cells were sorted on a BD Influx cell sorter. 2000 cells were sorted per well in a 96 well PCR plate (Eppendorf, 951020401) containing 10 µl of TCL buffer (Qiagen, 1031576) with 1% β-mercaptoethanol (Sigma, M3148). Following FACS the lysate was vigorously vortexed for 30 s, spun down, snap frozen on dry ice and stored at −80 °C until further processing.

RNA library preparation and sequencing: RNA sequencing libraries were prepared using a SMART-seq2 (Batch 1 and 2) and SMART-seq2-like (Batch 3 and 4)) protocol for cDNA preparation followed by Nextera XT DNA library preparation. Briefly, cDNA was prepared from 20-2000 FACS-sorted cells stored at -80°C and thawed to lyse. RNA was purified using a bead-based clean-up immediately followed by priming, reverse transcription with template switching using Thermo Scientific Maxima H Minus Reverse Transcriptase and Invitrogen Superase-In RNase Inhibitor, and enrichment with 18-22 cycles of PCR amplification. cDNA was evaluated using fluorescent-based assays including PicoGreen (Life Technologies), Qubit Fluorometer (Invitrogen), and Fragment Analyzer (Advanced Analytics). Final RNA libraries were prepared using the Nextera XT DNA Library Prep kit in accordance with the manufacturer's instructions. Briefly, 1 ng of amplified cDNA underwent tagmentation to fragment and tag cDNA with adapter sequences. Tagmented cDNA was then ligated with Illumina Nextera dual indices and enriched using 12 cycles of PCR amplification. Final libraries were evaluated using Picogreen (Life Technologies) and Fragment Analyzer (Advanced Analytics). For samples in the batch 1 and 2, libraries were pooled and sequenced on a HiSeq 2500 (Illumina) using 2 x 101bp (Batch 1) and 2 x 76bp cycles (Batch 2). For samples in batch 3 and 4, libraries were pooled and sequenced on a NovaSeq 6000 (Illumina) using 2 x 50bp cycles.

In ROSMAP, RNAseq gene expression measured from monocytes collected in 234 participants at the baseline. We obtained the raw FASTQ files by creating an account in AD Knowledge portal (ProjectSynID: syn2580853). We have processed the fastq files using the same HRS RNAseq QC pipeline that we used to process HRS RNAseq data and generated counts. In the ROSMAP study, we used RLE transformation to standardize the counts value and set prior.count = 6 to scale log 0 values to be positive.

3.2.6. RNAseq-based E-VAE classifier using a larger set of quality genes

We developed an initial classifier model to using all transcriptomics features that met the quality threshold. Among 48,956 genes in the HRS RNAseq dataset we selected 14,016 genes with log2cpm ≥ 3 in at least 10% of the study participants, of which 8127 genes were present in the ROSMAP study RNAseq data of 12,843 genes with quality gene expression threshold as used in the HRS. We used normalized gene expression levels of these genes in log2cpm and used Min-Max normalization to scale the distribution of all genes between 0-1 as input to the E-VAE classifier model.

3.2.7. RNAseq DEG-based E-VAE classifier model

We employed DGE analysis using 8127 genes to select features that were differentially expressed in participants with dementia in the HRS training set (N=204). We used various cut-offs to select features for the E-VAE classifier model based on multiple iterations evaluating the model performance and finally chosen differentially expressed genes with an FDR cut-off of < 0.5 as input to the RNAseq DGE-based E-VAE classifier. We normalized the selected features to the range between 0-1 using Min-Max normalization based on the input requirement of the VAE model.

3.3. Results

3.3.1. RNAseq-based E-VAE classifier model using whole genome-wide RNAseq data

We used 8127 genes that pass quality threshold of expression level as input for the RNAseq-based E-VAE classifier before implementing an association-based feature selection. After tuning hyper parameters, we finalized an encoder network with two hidden layers to generate the latent embedding of RNAseq data based on the model performance on HRS validation dataset.

3.3.1.1. Training and validation of the RNAseq-based E-VAE classifier model in the HRS

Table 2: Comparison of prediction performances between the RNASeq-based E-VAE classifier model and penalized logistic regression model in the HRS and ROSMAP study.

Models	Accuracy	Sensitivity/ Recall	Specificity	AUC
VAE model using HRS test (N=52)	0.69	0.73	0.65	0.65
VAE model using ROSMAP test (N=234)	0.63	0.75	0.63	0.70
Penalized LR model HRS test (N=52)	0.71	0.92	0.5	0.74
Penalized LR model ROSMAP test (N=234)	0.48	0.67	0.47	0.51

Table 2 shows the comparison of RNAseq-based E-VAE classifier model performance with a penalized logistic regression (LR) model using 8127 genes in the HRS validation data and the figure 4 shows the AUC-ROC curve. The RNAseq-based E-VAE classifier model achieved an accuracy of 0.69 with 95% CI [0.57, 0.82] and sensitivity of 0.73 in the HRS validation data. We found that the RNAseq-based E-VAE classifier had lower predictive accuracy but lower sensitivity in the HRS validation data compared the penalized LR model accuracy of 0.71 and sensitivity of 0.92.

3.3.1.2. Generalizability of the RNAseq-based E-VAE classifier model in the ROSMAP study

After tuning the hyper parameters as mentioned in Chapter 2, section 2.2.1.1, the parameters of the RNAseq-based E-VAE classifier model with highest accuracy in HRS validation data was fitted to the ROSMAP data to evaluate the generalizability of the model performance in an external dataset. The RNAseq-based E-VAE classifier model got an accuracy of 0.63 with 95% CI [0.57, 0.69] with an AUC of 0.70 in the ROSMAP data compared to the accuracy of 0.48 with an AUC of 0.51 using the penalized LR model as shown in figure 4. We found that the RNAseq-based E-VAE classifier model had higher generalizability in external dataset compared to traditional LR model.

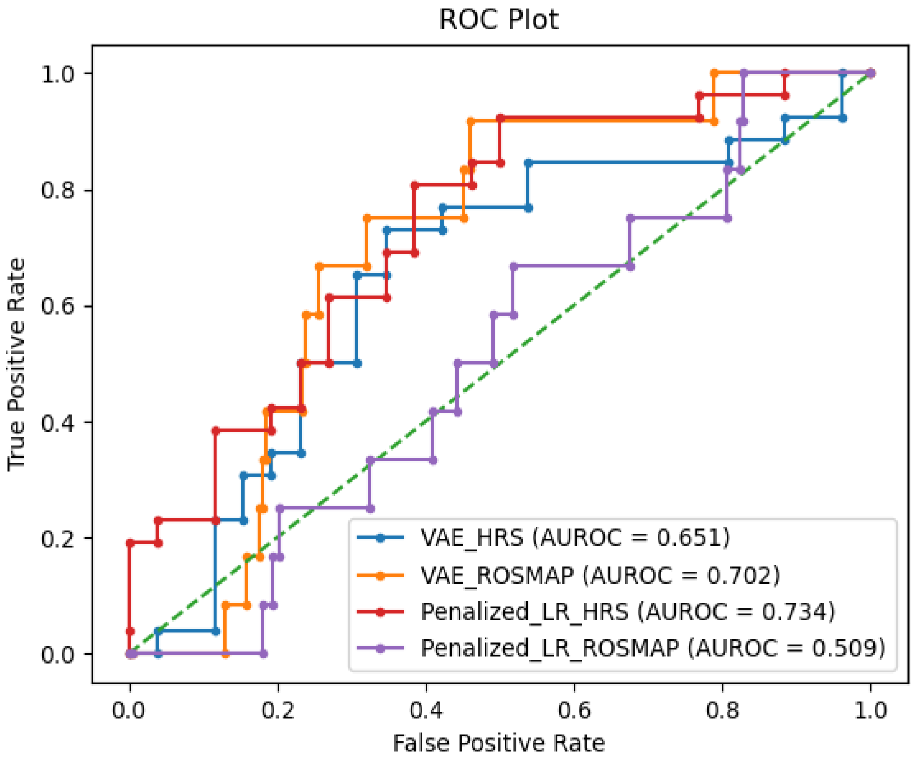


Figure 4: Comparison of AUC-ROC between the RNASeq-based E-VAE classifier model and penalized logistic regression model in the HRS and ROSMAP study.

3.3.2. RNAseq DGE-based E-VAE classifier model

We used an adjusted p value cut off of 0.5 to select 592 DEGs (upregulated = 50, downregulated = 542) as input to develop the RNAseq DGE-based E-VAE classifier model. After tuning hyper parameters, we finalized an encoder network with one hidden layer to generate the latent embedding of RNAseq DEGs based on the model performance on HRS validation dataset. Figure 5 shows the MA plot between the average log2CPM of the genes and the log2fold changes to visualize the distribution of DEGs.

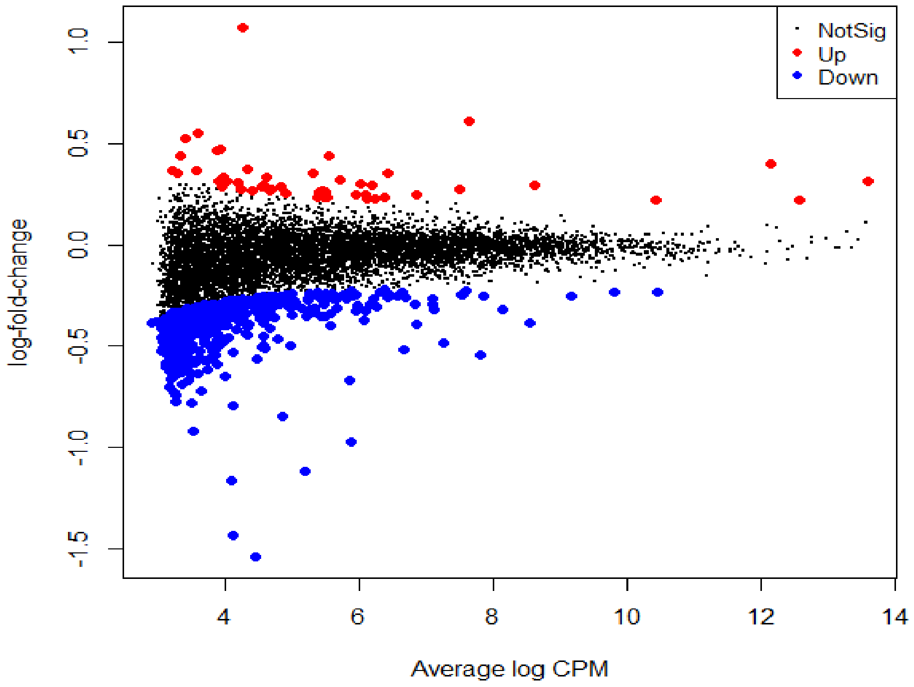
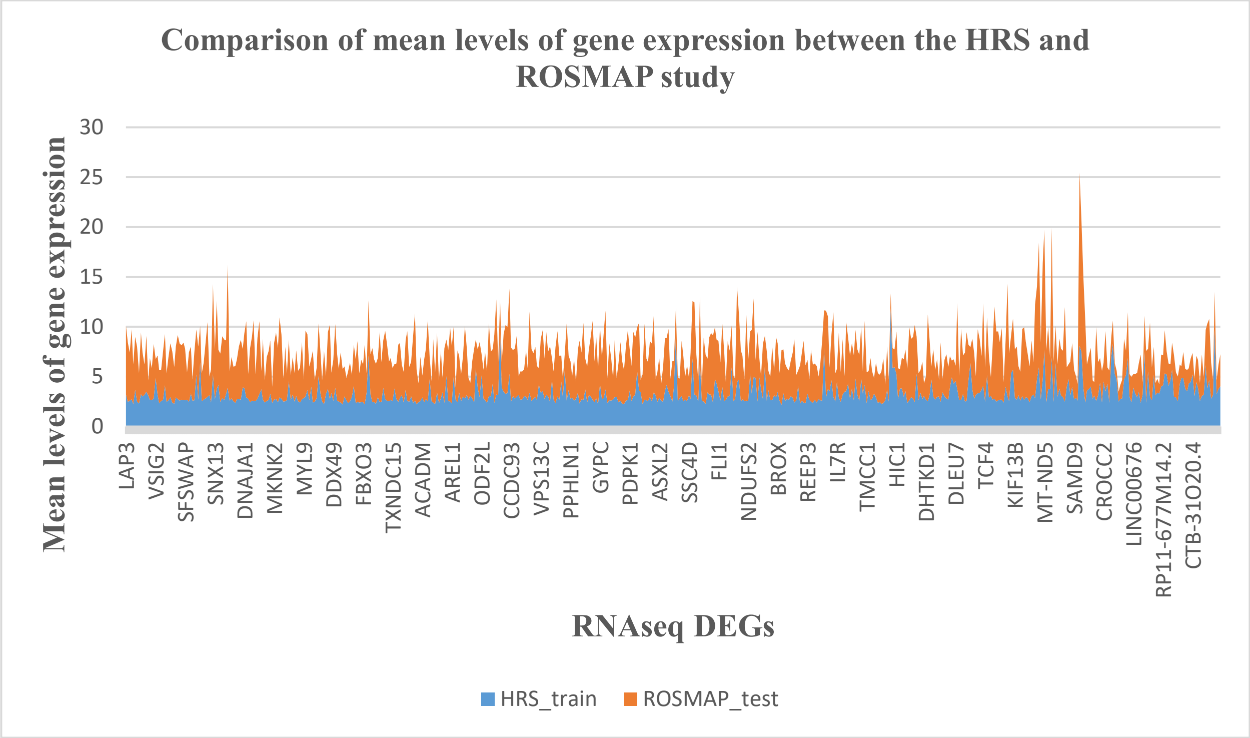


Figure 5: MA plot / Bland–Altman plot is the mean difference plot between log fold change and average normalized expression of the genes. The red dots represent upregulated genes and blue dots represent down regulated genes in the differential expression analysis at an adjusted p value of 0.5.

As shown in the supplementary figure 1, we observed higher levels of gene expression of DEGs in the ROSMAP study compared to the HRS.



Supplementary figure 1: Distribution of mean levels of gene expression by differentially expressed genes (DEGs) in the HRA train data and ROSMAP study test data.

3.3.2.1. Training and validation of the RNAseq DGE-based E-VAE classifier model in the HRS

Table 3 shows the comparison of RNAseq DGE-based E-VAE classifier model performance with a penalized logistic regression (LR) model using 592 differentially expressed genes in dementia in the HRS validation data and the figure 6 shows the AUC-ROC curve. The RNAseq DGE-based E-VAE classifier model achieved an accuracy of 0.83 with 95% CI [0.72, 0.93] and sensitivity of 0.77 in the HRS validation data. We found that the RNAseq-based E-VAE classifier had higher predictive accuracy but lower sensitivity in the HRS validation data compared the penalized LR model accuracy of 0.56 and sensitivity of 0.81.

Table 3: Comparison of prediction performances between the RNAseq DEG-based E-VAE classifier model and penalized logistic regression model in the HRS and ROSMAP study.

Models	**Accuracy**	**Sensitivity/ Recall**	**Specificity**	**AUC**
VAE model using HRS test (N=52)	0.83 (0.72, 0.93)	0.77 (0.61, 0.93)	0.88 (0.76, 1.00)	0.82
VAE model using ROSMAP test (N=234)	0.18 (0.13, 0.22)	0.92 (0.76, 1.00)	0.14 (0.09, 0.18)	0.60
Penalized LR model HRS test (N=52)	0.56 (0.42, 0.69)	0.81 (0.66, 0.96)	0.31 (0.13, 0.49)	0.62
Penalized LR model ROSMAP test (N=234)	0.58 (0.52, 0.64)	0.17 (0.00, 0.38)	0.60 (0.54, 0.67)	0.35

3.3.2.2. Generalizability of the RNAseq DGE-based E-VAE classifier model in the ROSMAP study

After tuning the hyper parameters, the parameters of the RNAseq DGE-based E-VAE classifier model with highest accuracy in the HRS validation data was fitted to the ROSMAP data to evaluate the generalizability of the association feature selection model in an external dataset. The RNAseq DGE-based E-VAE classifier model got an accuracy of 0.18 with 95% CI [0.13, 0.22] and sensitivity of 0.92 with an AUC of 0.60 in the ROSMAP data compared to the accuracy of 0.58 with an AUC of 0.35 using the penalized LR model. We found that the RNAseq DGE-based E-VAE classifier model had higher generalizability to classify dementia correctly in external dataset compared to traditional LR model.

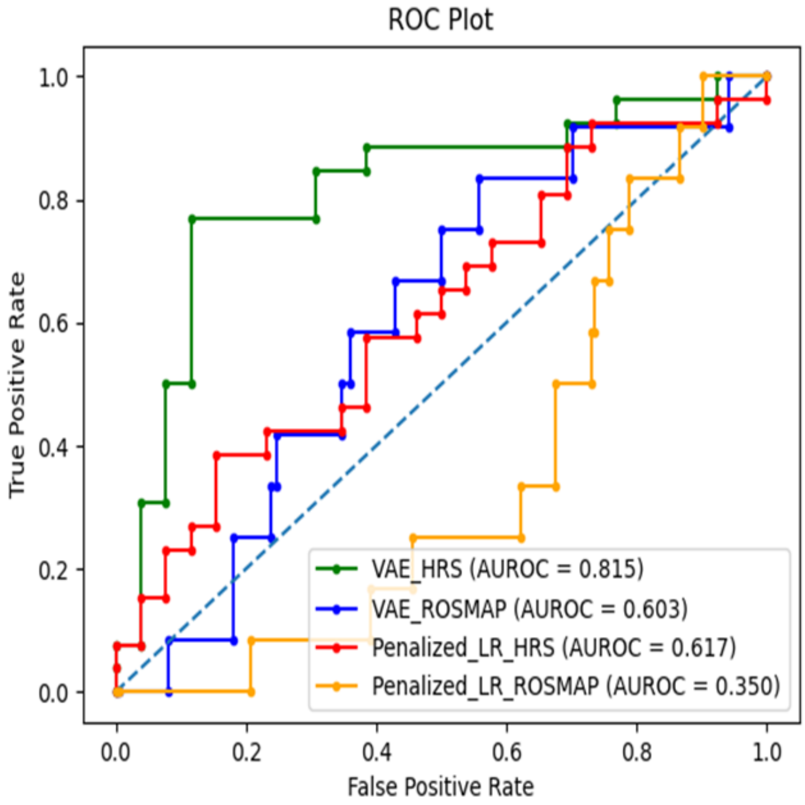


Figure 6: Comparison of AUC-ROC between the RNASeq DEG-based E-VAE classifier model and penalized logistic regression model in the HRS and ROSMAP study.

3.3.3. Comparison of performance between two RNAseq-based E-VAE classifier models

We observed that the RNAseq DGE-based E-VAE classifier model had high sensitivity and AUC in classifying dementia in the HRS validation data when compared to the RNAseq-based E-VAE classifier model. Additionally, the DGE-based E-VAE classifier model demonstrated greater sensitivity in identifying dementia in the ROSMAP study data compared to the RNAseq-based E-VAE classifier model. So we used only the latent features from the DGE-based E-VAE classifier model to evaluate the biological interpretability.

3.3.4. Biological interpretability of latent features from DEG-based E-VAE classifier model

We selected top-weighted 18 genes (Supplementary table 1) based 2.2 SD cut off for weights. The list of top weighted genes include many regulatory genes. None of the pathways were enriched in ORA, as shown in Figure 7.

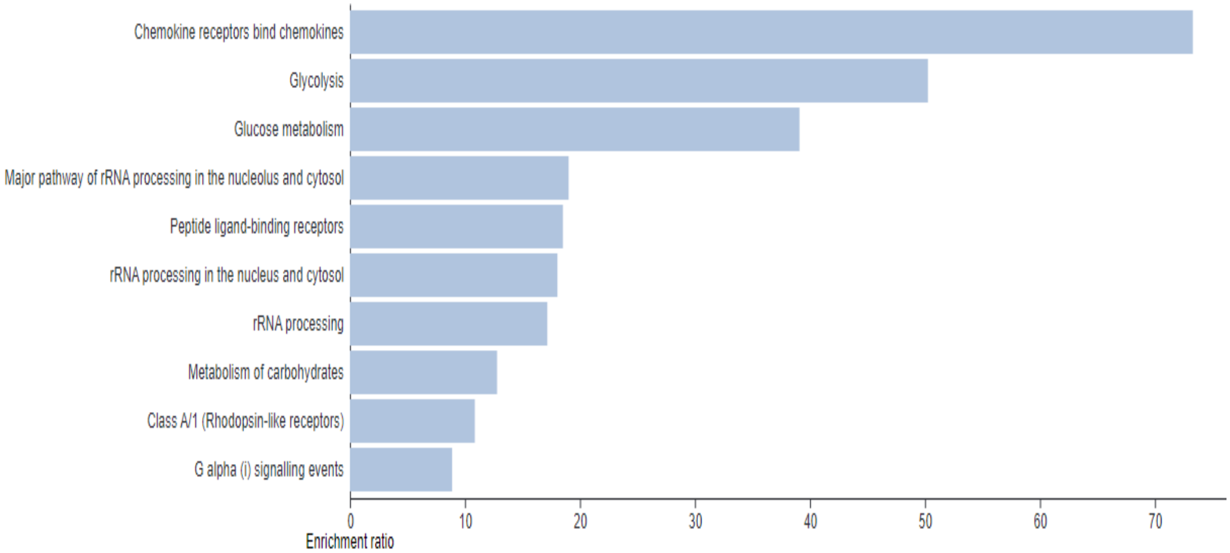


Figure 7: Results of over representation analysis (ORA) using Web based gene set analysis toolkit (WebGestalt).

This figure shows the pathway enrichment results of ORA using Reactome. We used the Benjamini-Hochberg (BH) method adjusted FDR p-value of 0.05 to select over enriched pathways.

3.4. Discussion

In this study, we demonstrated the application of E-VAE classifier model to incorporate whole genome-wide transcriptome data from blood to predict dementia among older adults in the HRS. We compared the performance of RNAseq-based E-VAE classifier model using a larger set of genes with a quality threshold level of gene expression and using differentially expressed genes in dementia. Our results showed that RNAseq-based E-VAE classifier performed better to classify dementia in an external data compared to standard regularized logistic regression model that suffered over fitting to the transcriptome profile in the HRS validation data. In addition, we found that RNAseq DGE-based E-VAE classifier model achieved high sensitivity to classify dementia in an external data compared to RNAseq-based E-VAE classifier model using whole transcriptome data. The top-weighted genes in the DGE-based E-VAE classifier included a mix of protein coding and regulatory genes.

In the last decade many studies have utilized machine learning methods to analyze RNAseq data from brain tissue samples to predict genes or interactions of genes [102] in the AD/ ADRD. Utilizing a joint density-based non-parametric Differential Interaction Network Analysis and Classification (JDINAC) [103] method using RNAseq data from the dorsolateral prefrontal cortex gray matter of the ROS and MAP studies identified 11 hub genes and 10 gene co-expression pairs. Their model achieved an AUC of 0.84 with accuracy of 0.79 and sensitivity of 0.78 [102]. The hub genes in their study were different from what we have observed, this is probably due to the difference in the input genes and expression levels of genes in brain and blood tissues. Multiple studies have analyzed gene expression data from NCBI Gene expression Omnibus (GEO) database of microarray studies and developed ML models to classify dementia. A recent study among 90 AD patients and 90 individuals without AD developed the prediction models using transcriptome data from microarray-based sequencing showed high performance of prediction models [97]. They mainly found that genes in the "mitochondrial translation" ontology were enriched in their list of DEGs, we also had mitochondrial ribosomal proteins (MRPs) associated genes (MRPL42 and MRPS26) in the list of DEGs in our study. They have implemented feature selection using PCA and DGE analysis on the whole dataset and then split the train and test data for model training and evaluation. This might have resulted in data leakage and thereby provided very high performance of their ML models. Another study in Malaysia among 92 AD patients and 92 non-AD subjects, gene expression profiling gathered was performed by microarray and validated through an independent gene expression approach using quantitative polymerase chain reaction (qPCR) [96]. Compared to our study they employed a random forest-based Boruta's algorithm for feature selection and achieved an accuracy of 0.82 and sensitivity of 0.85 using 16 gene transcripts in the internal validation data. Studies comparing the microarray and RNAseq methods showed that RNAseq-based gene expression shown to be more sensitive for differential gene expression compared to gene expression microarray for functional genomics applications [104, 105].

A recent large-scale RNA-seq transcriptome study on a large number of AD samples identified two top-ranked hub genes (*EEF2* and *RPL7*) as potential blood-based biomarkers for early diagnosis of AD with an AUC of 0.88 [98]. Compared to our study, the quality cut-offs used in their study was relaxed; we have discarded low-quality reads (< Q30) compared to their cut-off (< Q20) and we selected genes with threshold cpm ≥ 30 in at least 10% of study participants compared to their threshold cpm >1 in more than one fourth of all sequenced samples were used for DGE analysis. They identified the DEGs and hub genes using the whole sample and then developed the random forest classifier prediction models using a subset of all sample. This might led to the increased performance of their internal validation model based on clinical information (age, sex, and APOE ε4 genotypes) and three potential biomarkers obtained (i.e., the proportion of neutrophils, *EEF2*, and *RPL7*) and risk prediction model in the prospective cohort (accuracy = 0.73) [98]. We did not observe these hub genes in our study.

A study using large GTEx and TCGA tissue transcriptome data showed that transcriptomics-based phenotype prediction clearly benefits from proper normalization techniques and regularized regression approaches [106]. They showed that combining large number of genes outperformed single genes approaches for phenotype prediction. In their study, after evaluating multiple prediction models they found that deep representations from deep neural network-based architectures did not improved prediction compared to regularized linear models. However our study, by utilizing an end-to-end deep learning model, we have taken a two prong approach of disease classification and biomarker discovery. We demonstrated that deep representations learned using a VAE architecture along with a neural network-based classifier approach outperformed the regularized logistic regression models. In addition, adapting a linear decoder our explainable model identified biologically latent representations.

Compared to traditional linear methods used for transcriptome analysis, applications of advanced deep learning (DL) architectures can handle batch effects and non-linearity observed in high dimensional genome-wide transcriptome data to develop optimized prediction models [6, 7, 44]. Even though both the HRS and ROSMAP studies employed identical methods for RNAseq analysis, the HRS study used whole blood for whole genome-wide transcriptome analysis whereas the ROSMAP study used blood monocytes. This led to higher levels of gene expression in the ROSMAP study compared to the HRS (Supplementary figure 1), still the E-VAE classifier model we developed was able to successfully classify dementia in the ROSMAP study.

3.4.1. Strengths and limitations

Previous studies have not comprehensively evaluated the role of blood transcriptome to predict AD dementia in a large representative population of older adults. Availability of blood transcriptome from a large representative population of older adults allows for better generalizability of study results. While there have been numerous analytical approaches that have been utilized to analyze the blood transcriptome data, utilizing an end-to-end methods to analyze whole genome-wide transcriptome data from large epidemiological studies is an innovative approach. Compared to conventional ML methods, VAE-based methods uses a nonlinear technique to project high dimensional data to a low dimensional space. The biological interpretability of deep transcriptomic latent features generated from a deep learning model is another innovative approach. By translating the biological insights from the deep learning-based features may lead to a new biological hypothesis to be investigated. One major limitation of this study research is that it relied on cross-sectional data. The prediction model used to classify dementia was trained based on blood transcriptome data and outcome dementia collected at a single time point, which may not be suitable to predict incident dementia. Also relying solely on transcriptome data to predict dementia may not provide all the information captured in the genome necessary for accurate prediction of dementia. Considering an integrated multi-omics approach may improve the prediction of dementia.

Chapter 4: Explainable variational autoencoder (E-VAE) model using DNA methylation to predict dementia

4.1. Introduction

AD/ADRD has a complex pathology contributed by the environmental and genetic factors. DNA methylation (DNAm) is an epigenetic mechanism involving the transfer of a methyl group onto the C5 position of the cytosine to form 5-methylcytosine [107]. DNAm, which is associated with regulating gene transcription, can reveal significant epigenetic changes that are heritable [108, 109]. These modifications can affect gene expression and contribute to the development of diseases like Alzheimer's disease (AD). While genome-wide association studies (GWAS) have identified many genetic variants associated with AD, they may not capture the full extent of the complex interplay between genetic and environmental factors that contribute to the disease. By contrast, RNA Seq and DNA methylation analyses can provide a more complete picture of how environmental exposures and stressors over a person's lifetime can influence their risk of developing AD. Previous studies have identified DNA methylation marks in various brain region in association with AD. A recent preprint on peripheral blood DNA methylation analysis using bisulfite pyrosequencing of selected candidate genes *(NXN, TREML2* and *HOXA3*) showed that addition of DNA methylation marks along with plasma pTau181 levels improved the classification accuracy for AD [110]. Another study comparing the methylation levels of selected genes in blood based on prior knowledge between 120 late onset AD cases and 115 healthy controls found no difference in methylation levels between the groups [111]. Several studies including the HRS, Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Framingham Heart Study (FHS) Offspring cohorts reported an association between blood-based DNA methylation measures and dementia/ cognitive decline in older adults [112-114]. Another study evaluating the association between epigenetic clocks/ biological age acceleration and risk factors of AD reported stronger association between epigenetic age acceleration and behavioral and environmental risk factors of Alzheimer's disease (AD) compared to genetic risk factors [115]. In addition, a study in Lothian Birth Cohort 1921 showed found no association between DNA methylation-based measures of accelerated ageing and risk for dementia in the oldest-old [116]. A systematic review summarized findings of several EWAS studies investigated peripheral DNA methylation in AD. They reported that 32 of 48 studies they evaluated reported some association of peripheral blood DNA methylation and dementia but majority of them had limited sample size or constrained by the case-control status [117]. However, the epigenetic clocks do not provide biological insights into how DNA methylation based biomarkers affect dementia. Instead, they likely reflect a global aging process that also increases the risk of dementia and dementia is positively associated with age. To gain a better understanding of the complex mechanisms involved in Alzheimer's disease and related dementias (AD/ADRD), further studies employing a comprehensive set of DNA methylation features are needed, particularly in light of previous discrepancies observed in studies utilizing epigenetic clocks and epigenome-wide association studies (EWAs). Here we porpose to develop an E-VAE classifier model using differentially methylated CpG sites to classify dementia in the HRS.

4.2. Methods

4.2.1. DNAm-based E-VAE classifier model

We used DNAm data from 2714 participants in the HRS who had both genetic variants and RNAseq data available to develop the DNAm-based E-VAE classifier model. We utilized a similar architecture of E-VAE classifier model as described in Chapter 2 and 3. After tuning hyper parameters, we finalized an encoder network with one hidden layer to generate the latent embedding of DNAm features based on the model performance on HRS validation dataset.

DNA methylation: DNA methylation assays were performed on all study participants. The methylation profiles were generated by the Illumina Infinium Methylation EPIC BeadChip at the University of Minnesota Genomics Center that queries individual CpG sites to determine the % methylation. The 850,000-feature array is an update to Illumina’s popular HumanMethylation450 methylation platform. The beta values represent the methylation ratio of each CpG site with a range of 0-1. We removed 29431 probes from the total 866091 probes based on the minfi (R package) detection p-values threshold of 0.01 [118] along with the 168,615 flagged probes. High dimensional data with a small number of samples can lead to biased model performance estimates. To reduce the problems of having high dimensional data with small sample size and large number of methylation features, we employed association-based feature selection to input to the E-VAE classifier model.

4.2.2. Data preprocessing and differential methylation analysis

After the feature selection step, the missing values in the dataset were replaced by the mean of the corresponding methylation feature for the training and test data separately to generate a complete dataset. We used ChAMP analysis pipeline [119] for differential methylation analysis on the training dataset in the HRS and selected differentially methylated probes (DMPs) based on a p value threshold of 0.001 as input for the E-VAE classifier model.

4.3. Results

4.3.1. Training and validation of the E-VAE classifier model in the HRS

The details of 204 participants in the HRS training data is described in the Chapter 2. We chose 1451 DMPs in dementia compared to normal participants, from differential methylation analysis of 668,045 probes after removing flagged probes to develop the E-VAE classifier model. Table 4 shows the comparison of E-VAE classifier model performance with a penalized logistic regression (LR) model using 1451 DMPs in the HRS validation data and the figure 8 shows the AUC-ROC curve. The DMP-based E-VAE classifier model achieved an accuracy of 0.79 with 95% CI [0.68, 0.90] and sensitivity of 0.88 in the HRS validation data. The E-VAE classifier using DMPs had higher predictive accuracy but sensitivity in the HRS validation data compared the penalized LR model accuracy of 0.69 and sensitivity of 0.85.

Table 4: Comparison of prediction performances between the DMP-based E-VAE classifier model and penalized logistic regression model in the HRS study.

Models	Accuracy	Sensitivity/ Recall	Specificity	AUC
E-VAE classifier model using HRS test (N=52)	0.79	0.88	0.69	0.80
Penalized LR model HRS test (N=52)	0.69	0.85	0.54	0.75

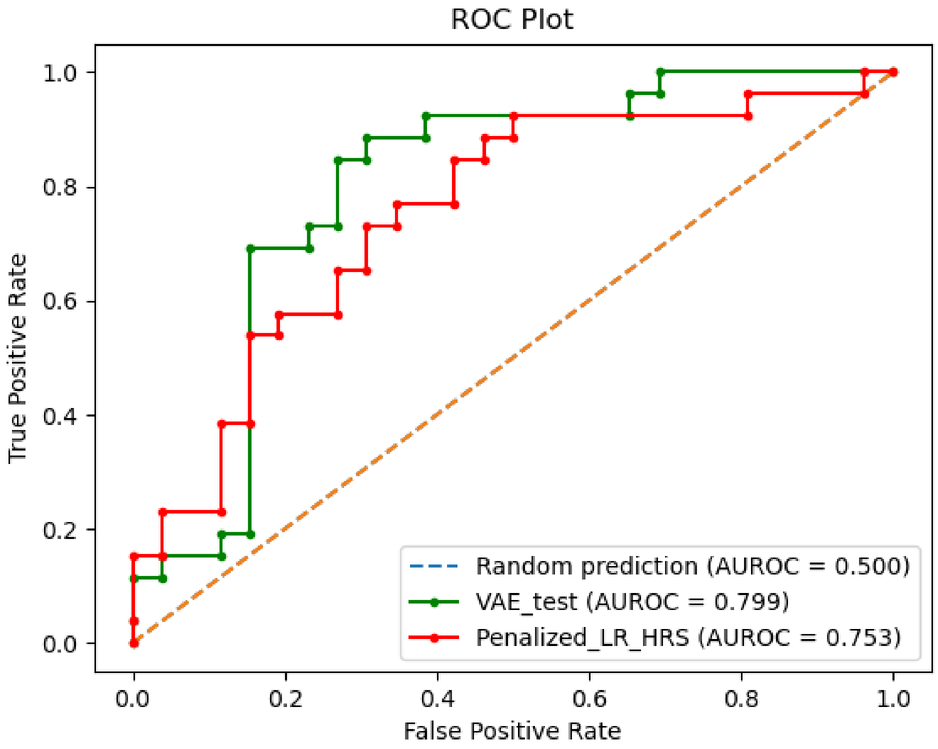


Figure 8: Comparison of AUC-ROC between the DMP-based E-VAE classifier model and penalized logistic regression model in the HRS.

4.3.2. Biological interpretability of the latent embedding of DMPs

We extracted 10 top-weighted DMPs that were 2.4 SD above or below the mean weight, the cut-off was chosen to limit the number of top weighted genes. Using the Illumina Methylation EPIC IDs and UCSC reference gene database, we mapped the top-weighted DMPs to 10 genes (Supplementary table 1) and used WebGestalt to characterize the top weighted genes. We did not observe any significant pathways over enriched by ORA using Reactome, shown in Figure 9.

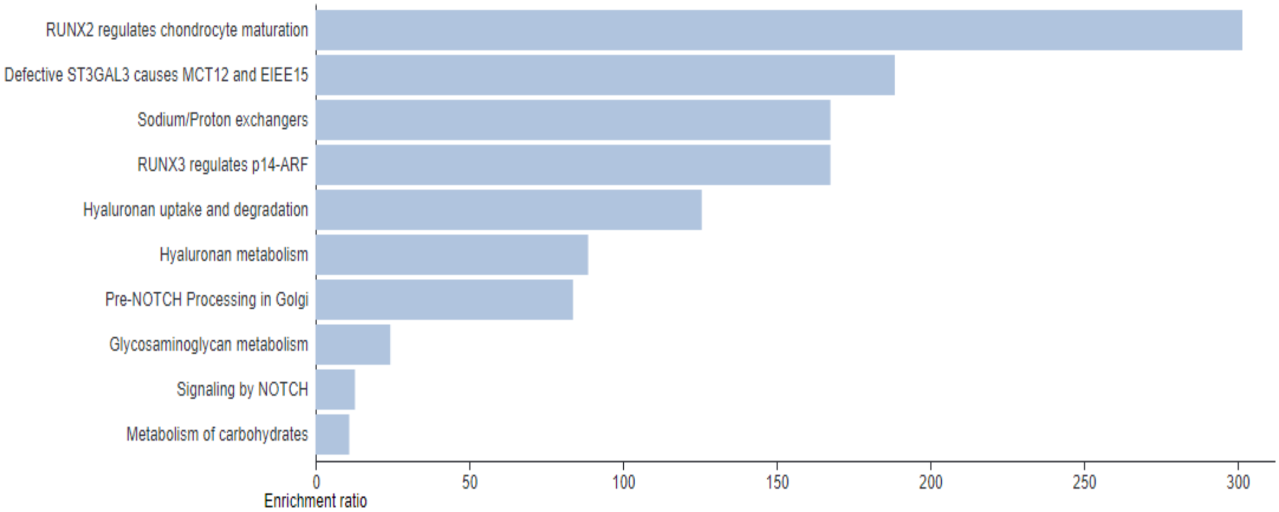


Figure 9: Results of over representation analysis (ORA) using Web based gene set analysis toolkit (WebGestalt).

This figure shows the pathway enrichment results of ORA using Reactome. We used the Benjamini-Hochberg (BH) method adjusted FDR p-value of 0.05 to select over enriched pathways.

4.4. Discussion

In this study, we used the E-VAE classifier model developed in Chapter 2 evaluate the predictive performance of the model using DMPs using a similar association based feature selection as we implemented the RNAseq DEG E-VAE classifier model. The E-VAE classifier model learned to generate latent embedding of high dimensional DMPs. The DMP-based E-VAE classifier model achieved higher accuracy (0.79) and AUC (0.80) compared to a penalized logistic regression model (accuracy = 0.69, AUC = 0.75). Using the explainable prediction model we extracted the input feature weights and evaluated the biological interpretability of low dimensional latent embedding of DMPs. Compared to the single omics model implemented in Chapter 2 and 3, a major limitation of this study is the lack of external dataset for validation of the generalizability of the DMP-based E-VAE classifier model.

We selected differentially methylated CpG sites from the comprehensive methylation profile in blood. As described previously, features selected using association with dementia improved the performance of DNAm DMP-based E-VAE classifier model. Recently a study using ADNI DNA methylation data in blood showed that deep autoencoders outperform the state of the art ML models to predict the AD progression [120]. They performed a feature selection based on the ratio of two variances of beta values in conversion group (e.g. CN-to-MCI) and non-conversion group (e.g. CN-to-CN). Their models achieved an AUC of 0.99 for predicting cognitive normal to MCI progression and achieved an AUC of 0.77 for predicting MCI to AD progression. Compared to our model, they also achieved best performance using 1000 most informative CpGs. We did not observe the CpG site 'cg07847171' that they found near to PSEN2 gene. However we found many biological processes common to their study analysis such as neurogenesis, neuron and nervous system development and signal transduction.

Chapter 5: Explainable variational autoencoder (E-VAE) classifier model using multi-omics data to predict dementia

5.1. Introduction

Alzheimer's disease (AD) and AD related dementia (ADRD) are complex progressive neurodegenerative diseases characterized by memory loss and cognitive decline leading to dementia among older adults. Studies have shown that dysfunctional interactions of genes, gene expression and biochemical changes occur many years before the onset of clinical symptoms of AD [1, 2, 12]. Recent advancement in high throughput technologies have generated vast amount of molecular data including genetic variants, epigenetic, transcriptomics, proteomics and metabolomics data from blood and brain tissues improved our understanding of complex molecular mechanisms associated with pathways of ADRD [3]. While genome-wide association studies (GWAS) have identified several genetic risk factors for AD, they cannot fully capture the complexity of the disease on their own. On the other hand, RNA sequencing (RNASeq) can provide information about gene expression and cellular function, which can be influenced by both genetic and environmental factors. By combining data from multiple omics, we can gain a more holistic view of the disease and identify novel biomarkers, therapeutic targets, and potential environmental interventions. For example, a multi-omics approach can identify genetic risk factors that influence the expression of specific genes or pathways, as well as environmental factors that modulate gene expression. By integrating data from multiple omics, we can also identify gene regulatory networks and pathways that are dysregulated in AD, providing insights into the underlying biological mechanisms of the disease compared to single omics studies [4].

The use of multi-omics integration techniques in translational medicine is still evolving with the recent availability of complex high dimensional molecular data from blood samples in large population studies such as HRS [35] and NIH initiatives such as Trans-Omics for Precision Medicine (TOPMed) program [20]. So it is necessary to develop methods to integrate multi-omics data. We hypothesized that a holistic approach of integrating hierarchical multi-omics data may shed light on cross-talk patterns between various functional omics and allow us to understand the new biology in AD related dementias. So we expanded our single-omics based E-VAE classifier model to include GWAS SNPs and DEGs in an integrated method to develop a multi-omics based E-VAE classifier model for dementia in the HRS. We utilized the basic architecture of data-driven unsupervised modeling of VAE to extract the biologically relevant latent space from the multi-omics E-VAE classifier model.

Traditionally parametric approaches such as penalized regression models or canonical correlation analysis has been used to analyze multi-omics data. These approaches are based on the assumption that the data follows a specific distribution and that the relationship between features is linear. In case of biological systems, especially when studying phenotype such as AD/ADRD with complex and heterogeneous biological processes, these assumptions may not hold well where the relationships are non-linear and complex. Moreover, these approaches may not be able to capture the interactions and dependencies among the omics features, which are critical for understanding complex biological processes. In contrast, deep learning models can automatically learn the complex relationships and patterns in the data without making any assumptions about the distribution or linearity of the data. These models can handle high-dimensional and multi-modal data, such as multi-omics data, and can identify the interactions and dependencies among the omics features. Deep learning methods have been applied successfully using heterogeneous multi-omics data from blood with high accuracy for classifications of various types of cancers and brain diseases [6, 7]. Thus, newer deep learning methods applied to multi-omics data can provide a more powerful and comprehensive approach to elucidating the biological mechanisms underlying complex diseases such as AD/ADRD and biological processes compared to traditional parametric approaches.

Here we utilized the Variational Autoencoder (VAE) architecture proposed by Kingma and Welling [47] that can handle high order interactions between and within high dimensional omics data that impose a set of distributions on the latent variables to regularize latent space and enhance the classification performance compared to traditional autoencoders. As described previously in Chapter 2, the proposed deep generative neural network prediction model, the E-VAE classifier model has a basic structure of VAE (with a fully connected encoder, hidden layers and a linear decoder) which enabled us to integrate multi-omics data without a prior hypothesis and a classifier network for supervised prediction of dementia. In our previous studies we showed that using genetic variants and differentially expressed genes (DEGs) in single omics based explainable variational autoencoder (E-VAE) classifier can handle higher order interactions within omics data and demonstrated the generalizability of the E-VAE classifier model. Here we expand the single-omics based E-VAE classifier model to integrate multi-omics data to classify dementia and to extract biologically relevant features from latent embedding of multi-omics data.

We used genetic variants and transcriptomics data from the Health and Retirement Study (HRS) to develop and validate the multi-omics E-VAE classifier model to predict dementia. We validated the generalizability of the developed model in an external dataset from the Religious Orders Study and Rush Memory and Aging Project (ROSMAP). We also demonstrated the biological interpretability of learned latent representation from the multi-omics features using a linear decoder approach along with optimizing the classifier network to predict dementia. This is the first study utilizing a deep learning neural network model to integrate multi-omics data to predict dementia and demonstrating the external validity in an independent study.

5.2. Methods

5.2.1. The architecture of multi-omics E-VAE classifier model and its optimization

As described in Chapter 2, we adapted a basic architecture of VAE classifier network from the OmiVAE model and implemented a linear decoder to extract the weights associated with latent representations of input multi-omics data. Compared to the single–omics E-VAE classifier models, the multi-omics E-VAE classifier model generate latent representation from GWAS SNPs and DEGs together by optimizing the classifier network for dementia. The deep neural network has multiple hidden layers with each layer successively refines the results of the previous layer in unsupervised learning. Compared to other auto encoders (AEs), VAE has a more regularized encodings distribution to avoid overfitting and the specific properties of latent space enable new sample generation/data reconstruction. [47].

We have evaluated various VAE architectures to integrate the features from genetic and transcriptomic data and selected the architecture that gave the best performance in the HRS validation data. As shown in figure 10, we used 2 hidden layers for GWAS SNPs and a hidden layer for DEGs, then we added another hidden layer to integrate the two omics features before the latent layer. The other encoder networks we experimented were: one hidden layer for each omics data, two hidden layers for each omics data, one hidden layer for DEGs and two hidden layers for GWAS SNPs. We achieved the highest performance in the HRS validation data using hidden layers to individually train the omics data to reduce the dimensions and then used a hidden layer to integrate the two omics data before generating the latent space. Then used the linear decoder learn to generate reconstructed GWAS SNPs and DEGs by optimizing the unsupervised phase. We used a supervised classifier network with two hidden layers with a final layer of 2 dimensions to classify the outcome dementia. The classification network added additional regularization for learning the latent representation from omics data that enable the model to learn latent features important to classify participants with dementia from normal. As described in single-omics studies, the joint loss function of the E-VAE classifier model was a combination of VAE loss and classification loss. We considered the transcriptomics features as x_g and reconstructed feature as $x`_g$ and genetic variants as $x_v$ reconstructed feature as $x`_v$. So the loss function of unsupervised VAE was L_{VAE} =cross-entropy (x_e,$x`_e$) + cross-entropy ($x_v$,$x`_v$) + L_{KLD} .The total loss function of the multi-omics E-VAE classifier model was $L_{TOTAL} = L_{VAE} + L_{Dementia}$; where $L_{Dementia}$ is the binary cross-entropy loss.

We used the Adam optimizer to the train the multi-omics E-VAE classifier model and used Rectified Linear Units (ReLU) activation function and batch normalization to each fully connected block in the encoder hidden layers, a sigmoid activation was used in the linear decoder and a softmax function employed in the final layer of classifier. We used Optuna, a hyperparameter optimization framework to tune the hyperparameters (such as batch size, learning rate, and number of hidden layers for each omics, hidden layers and latent layer dimensions) in the multi-omics E-VAE classifier model. We evaluated the performance of the multi-omics E-VAE classifier model based on the total accuracy score and sensitivity

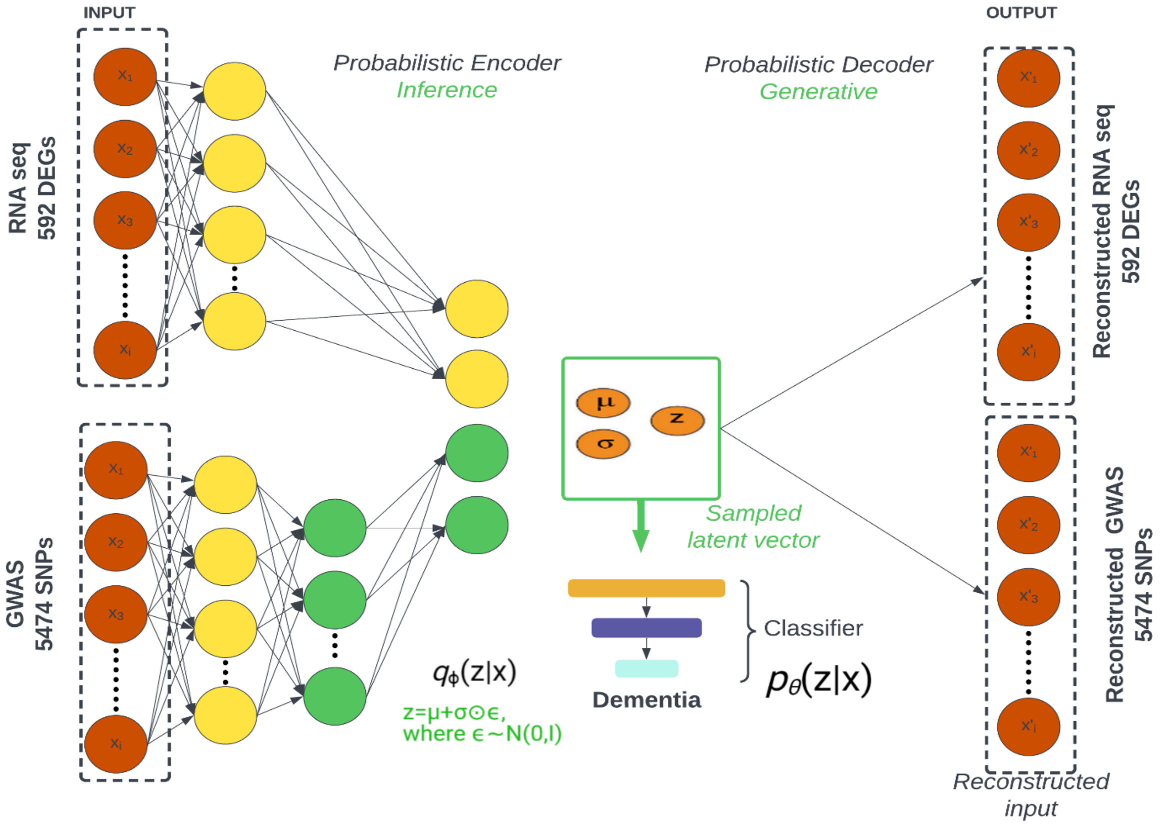


Figure 10 Illustration of the multi-omics E-VAE classifier model to predict dementia: This diagram shows a conceptual illustration of the study design by integrating selected features from genetic, and transcriptomic data as input features to the multi-omics E-VAE classifier model in the unsupervised phase and then latent representation from bottleneck layer (z) was used as input for the supervised classifier network. The linear decoder learned to generate reconstructed genetic and transcriptomic data from encoded latent features.

to classify participants with dementia. We also used a Receiver-operating characteristic (ROC) analysis to evaluate the non-random prediction performance of the models and used the area under the curve (AUC) to compare the performance of various prediction models [75]. The model evaluation metrics and implementation were same as in Chapter 2.

5.2.2. Comparison of VAE model with penalized logistic regression (LR) models

We used a penalized logistic regression model as our base model to evaluate the performance of multi-omics model compared to the single-omics model. We concatenated the GWAS SNPs and DEGs datasets and used as input to the ElasticNet model using scikit-learn 1.2.1 library and used that to compare the performance of the multi-omics E-VAE classifier model. We optimized the ElasticNet model by tuning the hyper parameters such as l1_ratio parameter (penalty type) for a model regularizer and inverse of regularization strength.

5.2.3. Biological interpretation of latent representation from integrated multi-omics data

We used a similar method as described in Chapter 2 to select the top weighted features from the multi-omics E-VAE classifier model. We selected top weighted SNPs with absolute z-score value for weight ≥ 2.3 and RNAseq DEGs with absolute z-score value for weight ≥ 2.1 to limit the number of features for biological interpretation. We utilized the WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) to characterize the top weighted genes from the GWAS SNPs and DEGs and performed Network Topology Analysis (NTA) to translate the top weighted gene list into biological insights. We also utilized NCBI gene database to curate the top weighted genes biological processes and evaluated the set of top-weighted genes with ALZgene [121] top results. The AlzGene database is a collection of published Alzheimer's disease genetic association studies, with random-effects meta-analyses for polymorphisms with genotype data in at least three case-control samples. In addition, we have evaluated the top ranking neighbors in the NTA in the context of AD/ADRD.

5.2.4. Data

As mentioned in Chapter 2 in detail, the HRS included measures of cognitive functioning to understand the onset and impact of cognitive impairment using an in person and telephonic surveys [76]. We utilized the genome-wide transcriptomic, and genetic variants (GWAS SNPs) data along with cognitive function scores measured in 2016 survey to build the multi-omics E-VAE classifier model and utilized the ROSMAP study data to validate the generalizability of the model developed in the HRS.

Ascertainment of outcome dementia: We harmonized the cognition measures and definition of dementia in the HRS and ROSMAP study using z-scores for cognition tests that were available in both studies as described in Chapters 2.

HRS training and internal validation data: Among the 2714 study participants who had both genetic variants and transcriptome data available, only 5% (N=128) had dementia in 2016. We used a balanced class approach by randomly selecting 128 participants from the 'Normal' class and the 128 participants with dementia to develop the multi-omics E-VAE classifier model. Then we split the data for training and validation using an 80:20 stratified split. The training data included 204 participants and internal validation data included 52 participants.

5.2.5. Multi-omics data processing in the HRS and ROSMAP study

HRS and ROSMAP study GWAS SNPs: We included 5474 genome-wide significant SNPs (p<5e-08) curated in the HRS based on already published large-scale GWAS studies related to ADRD and that were common in the ROSMAP study imputed dosage data. We used the same set of GWAS SNPs that we used in single omics E-VAE classifier model and additional details of GWAS SNPs pre-processing can be found Chapter 2.

HRS and ROSMAP study DEGs: As described in Chapter 3, we included 592 DEGs in dementia based on the differential expression analysis the HRS training data using edgeR based on FDR adjusted p-value cut off of 0.5 that provided the best performing single omics E-VAE classifier model in the HRS validation data as described in Chapter 3. All these genes were present in the ROSMAP study data but at a higher level of expression as shown in the supplementary figure 1 in Chapter 3. We used same set of DEGs that we used in single omics E-VAE classifier model and additional details of data pre-processing, quality analysis and feature methods were described previously in Chapter 3.

5.3. Results

5.3.1. Training and internal validation of the E-VAE classifier model in the HRS

We have included 102 dementia cases and 102 normal participants in the training set with an average age of 73 years, 110 women and 113 Whites and 26 dementia cases and 26 normal participants in the validation set with an average age of 72 years, 34 women and 29 Whites as mentioned in Chapter 2. Table 5 and figure 11 shows the performance of multi-omics E-VAE classifier model using 5474 SNPs and 592 DEGs compared to penalized logistic regression model. The multi-omics E-VAE classifier model achieved an

Table 5: Comparison of prediction performances between the multi-omics-based E-VAE classifier model and penalized logistic regression model in the HRS and ROSMAP study.

Models	**Accuracy**	**Sensitivity/ Recall**	**Specificity**	**AUC**
E-VAE classifier model using HRS test (N=52)	0.73 (0.61, 0.85)	0.73 (0.56, 0.90)	0.73 (0.56, 0.90)	0.80
E-VAE classifier model using ROSMAP test (N=234)	0.58 (0.51, 0.64)	0.75 (0.51, 0.995)	0.57 (0.50, 0.63)	0.66
Penalized LR model HRS test (N=52)	0.67 (0.55, 0.80)	0.77 (0.61, 0.93)	0.58 (0.39, 0.77)	0.72
Penalized LR model ROSMAP test (N=234)	0.73 (0.67, 0.79)	0.33 (0.07, 0.60)	0.75 (0.70, 0.81)	0.56

accuracy of 0.73 with 95% CI [0.61, 0.85] and sensitivity of 0.73 with an AUC of 0.80 in the HRS validation data. We found that the multi-omics E-VAE classifier had similar predictive accuracy and higher AUC but lower sensitivity in the HRS validation data compared the penalized LR model accuracy of 0.67, AUC of 0.72 and sensitivity of 0.77.

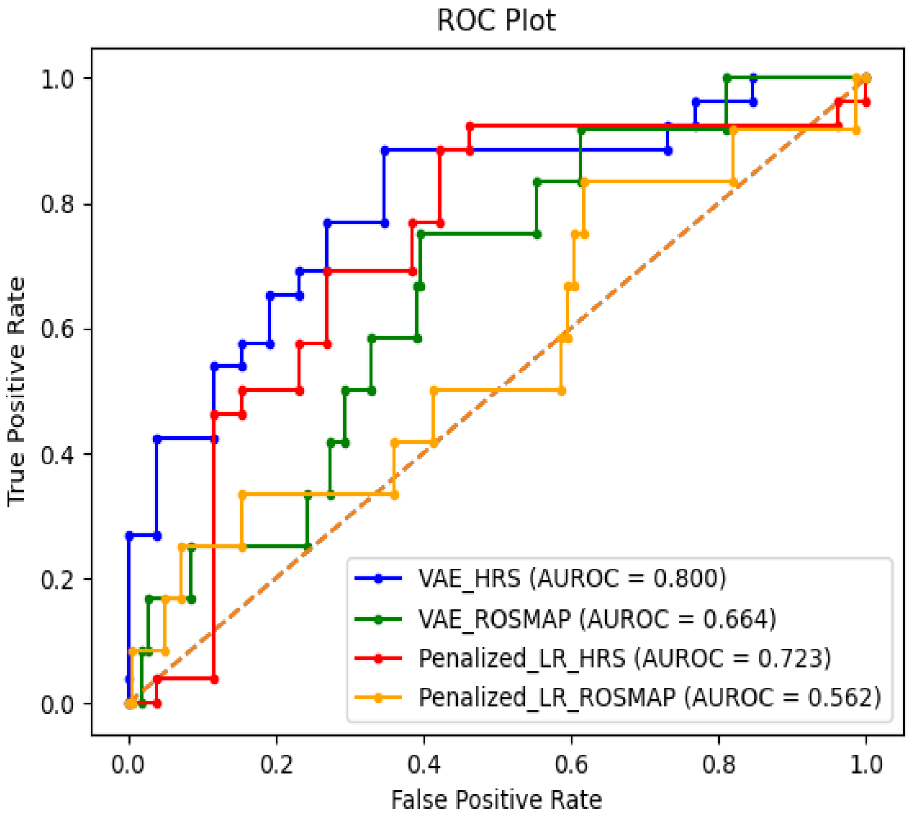


Figure 11: Comparison of AUC-ROC between the multi-omics E-VAE classifier model and penalized logistic regression model in the HRS and ROSMAP study.

5.3.2. External validation of the multi-omics E-VAE classifier model in the ROSMAP study

We have included 234 participants in the ROSMAP study who had both genetic variants and RNAseq data available. We have included 12 (5%) dementia cases and 222 (95%) normal participants in the external validation set with an average age of 84 years, included all Whites and 170 women**.** The parameters of the multi-omics E-VAE classifier model with highest accuracy in HRS validation data was fitted to the ROSMAP data to evaluate the generalizability of the model performance in an independent dataset. The multi-omics E-VAE classifier model achieved an accuracy of 0.58 with 95% CI [0.52, 0.64] and sensitivity of 0.75 with an AUC of 0.66 in the ROSMAP data compared to the accuracy of 0.73 and sensitivity of 0.33 with an AUC of 0.56 using the penalized LR model as shown in table 5 and figure 11. We found that the multi-omics E-VAE classifier model had higher generalizability in external dataset compared to traditional LR model.

5.3.3. Biological interpretability of the latent representation from integrated multi-omics of SNPs and DEGs

The top weighted genes extracted from the multi-omics E-VAE classifier model as shown in supplementary table 2 were used for biological interpretability of latent embedding. The figure 3 shows top ranked neighbors highlighted in blue from the network topology analysis that had reported interactions with the input top-weighted seed genes. The 10 top-ranked neighbors in the NTA analysis in figure 12 included several genes that were extensively studied in AD/ADRD were *RHOU, AKAP4, MCM2, APP, TRIM25, RNF4, NTRK1, HNRNPL and NUFIP*1. The analysis of biological processes involved by the top-weighted genes are listed in Supplementary table 3. The biological processes represented by top-weighted features extracted from both omics data includes Apoptotic process , cell division and cell surface receptor signaling pathway, neuron growth and development, protein transport, regulation of NF-kappaB signaling, positive regulation of cell migration and cell proliferation, integrin mediated pathway.

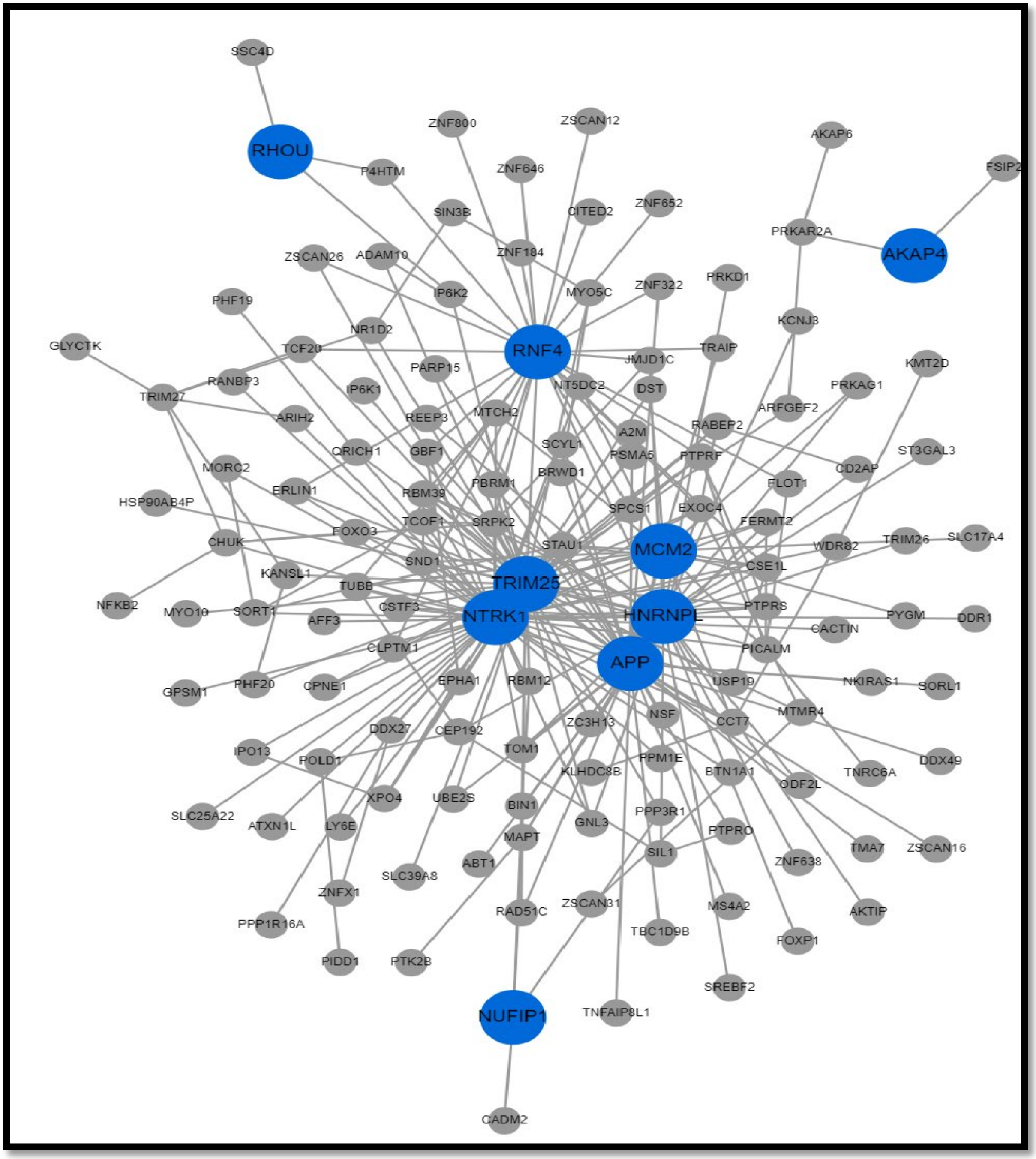


Figure 12: Network topology analysis (NTA) of top-weighted genes from E-VAE classifier using Web based gene set analysis toolkit (WebGestalt).

This figure shows the top-weighted genes from multi-omics E-VAE classifier model as seed genes and top-ranked neighbors of the genes highlighted in blue in PPI network.

5.4. Discussion

We developed an end to end deep learning model with an unsupervised phase to integrate multi-omics data and a supervised classifier network that achieved improved prediction for dementia compared to our published single omics-based studies to predict dementia. The multi-omics E-VAE classifier model achieved an accuracy of 0.73 with an AUC of 0.80 in the HRS validation data and achieved an accuracy of 0.67 with an AUC of 0.66 in the ROSMAP study data. We found that our model had better generalizability in an external dataset compared to penalized logistic regression model. We also demonstrated the utility of a linear decoder to extract the input feature weights for the latent embedding of integrated SNPs and DEGs for the biological interpretability by optimizing a classifier model for dementia. We found that the top weighted genes included a list of genes that were extensively studied in AD/ neurodegeneration and some novel genes that are in the pathways studied for AD/ADRD but not reported an association with AD/ADRD. Biological systems modeled using high throughput next-generation sequencing data from various facets and biological interactions are necessary to understand complex molecular mechanisms in AD dementia. Genome-wide association studies (GWAS) of AD have identified many new genetic loci [26, 27], affecting pathways of immunity, lipid metabolism, tau binding proteins, glucose metabolism, mitochondrial function [28] and amyloid precursor protein (APP) metabolism [29, 30]. An integrated multi-omics (metabolomics and trasncriptomics) study (N=78; 41 amyloid (+) and 37 amyloid (−)) in the INSIGHT-preAD cohort of elderly asymptomatic individuals showed that the ten top metabolites and transcripts analyzed using (sparse partial least squares-discriminant analysis) represented the most discriminant omics features with 99·4% chance prediction for amyloid positivity using sparse partial least squares-discriminant (sPLS-DA, a case of sparse generalized canonical correlation analysis (sGCCA)) and principal component analyses [39]. In concordance with our study findings, they also demonstrated that multi-omics integrative analyses better discriminated individuals with amyloid status than single omics analysis. An integrated omics network analysis using genomics, plasma and cerebrospinal fluid (CSF) metabolomics and AD risk factors found that imputed gene expression levels in blood were indirectly linked to AD risk factors through metabolites, 38 of 364 CSF metabolites explained 64% of variance observed in CSF p-tau [34]. A Convolutional Neural Network based model using electronic health records, imaging and SNPs data in ADNI study (N=220) showed that deep learning models outperforms shallow models for classifications of AD from cognitive normal [122]. Another deep neural network- based prediction model using gene expression and DNA methylation data from brain tissues using association based feature selection achieved an accuracy of 0.83 in the test dataset [123]. This study found only one gene, *MS4A4A*, from their list of selected genes overlapped with Alzgene database, whereas we got 5 genes (*BIN1, CR1, PICALM, MS46A* and *MS4A4E*) from our top-weighted genes match with top 10 results in AlzGene. The top Alzgenes not present in our top-weighted gene list were either had low expression in the RNAseq data or not identified as differentially expressed in blood among dementia vs. control samples in our study data. In a hospital-based brain aging study (N=120), using various omics measured in CSF and utilizing multi-omics factor analysis (MOFA) demonstrated that adding analytes selected by the MOFA model to APOE status can improve the prediction of cerebral AD pathology and cognitive decline [124]. All these studies had a small sample size (N = 70 – 220) compared to our study and did not validate the generalizability of the deep learning models in an independent study. Though studies in brain tissue and CSF are very useful for identifying novel biological determinants of AD, they remain of limited value in population-based screening efforts for the early identification of AD. For such efforts, biomarkers obtained from blood that can be easily obtained from potential patients are essential. However, there is very limited data on the utility of blood-based biomarkers in the early diagnosis of AD. A small study (N=80) analyzing untargeted metabolomics and proteomics revealed new biomarkers in AD and showed that integrated analysis allowed them to capture overall changes in biochemical pathways associated with MCI and AD [125]. A recent study evaluating the multi-omics (CSF and plasma proteomics, plasma metabolomics and SNP genotyping for AD PRS) relating to

AD endotypes based on A/T/N framework using correlation network analysis found individual proteins/metabolites and networks that have causal link with AD [126].

Previous studies demonstrated that single omics analyses using traditional models are not sufficient to characterize a trait and often lead to overfitting [36]. We demonstrated that deep learning-based E-VE classifier model is more generalizable than penalized logistic regression models to predict dementia using integrated multi-omics. Previous studies have not comprehensively evaluated the role of blood-based multi-omics data that include whole-genome wide genetic variants and transcriptomic data to predict dementia in a large representative population of older adults. The framework we developed allow us to integrate multi-omics features from genetic and transcriptomic data to evaluate the contribution of features from single omics studies to predict dementia in a multi-omics feature space. We demonstrated the utility of linear decoder approach to extract top-weighted features from each omics that contribute to the latent features and evaluated the biological meaning of the learned latent features from multi-omics feature space in dementia prediction. Recently, there have been studies that utilized ultrasensitive blood immunoassay to measure AD protein biomarkers, such as p-tau 231 [87] and p-tau181 [127], demonstrated a high level of accuracy in discriminating AD/Aβ+ from controls/Aβ-, with an AUC range of 0.90 – 0.994 in both discovery and validation cohorts. Notably, these studies using single protein biomarkers achieved a higher level of prediction performance than the multi-omics model that was developed. However, the objective of our study was not solely to develop a prediction model; rather, we aimed to translate the biological insights obtained from the latent embedding of the multi-omics E-VAE classifier model. This approach enabled us to comprehend the concealed biology in complex AD biological mechanisms through the utilization of data from multiple levels of biological regulation.

5.4.1. Biological interpretability of latent embedding of multi-omics data (DEGs and GWAS SNPs)

Studies showed that an integrated multi-omics approach provides additional insights regarding the flow of information underlying a disease mechanism compared to a single omics analysis [4, 36]. Recent advancements in deep learning methods allowed us to analyze high throughput data generated from omics studies in AD [84] to identify novel biomarkers for early detection. We showed that an integrated analysis of omics data provided a comprehensive view of different aspects of biological phenotype from multi-level omics data. We have extracted 20 top-weighted genes from the GWAS SNPs and 31 top-weighted genes from RNAseq DEG feature list input to the multi-omics E-VAE classifier model. We observed that the extracted top-weighted genes from GWAS SNPs or DEGs included genes that share common biological processes (for ex: SNPs on DCC and PIDD1 gene involved in apoptotic process) as well as genes that are involved in discrete biological processes (for ex: SNPs on ADAM10 is involved in amyloid precursor protein catabolic process and gene TCOF1 is involved in neural crest cell development).

5.4.1. Comparison of top weighted features and biological processes between single omics and multi omics E-VAE classifier models

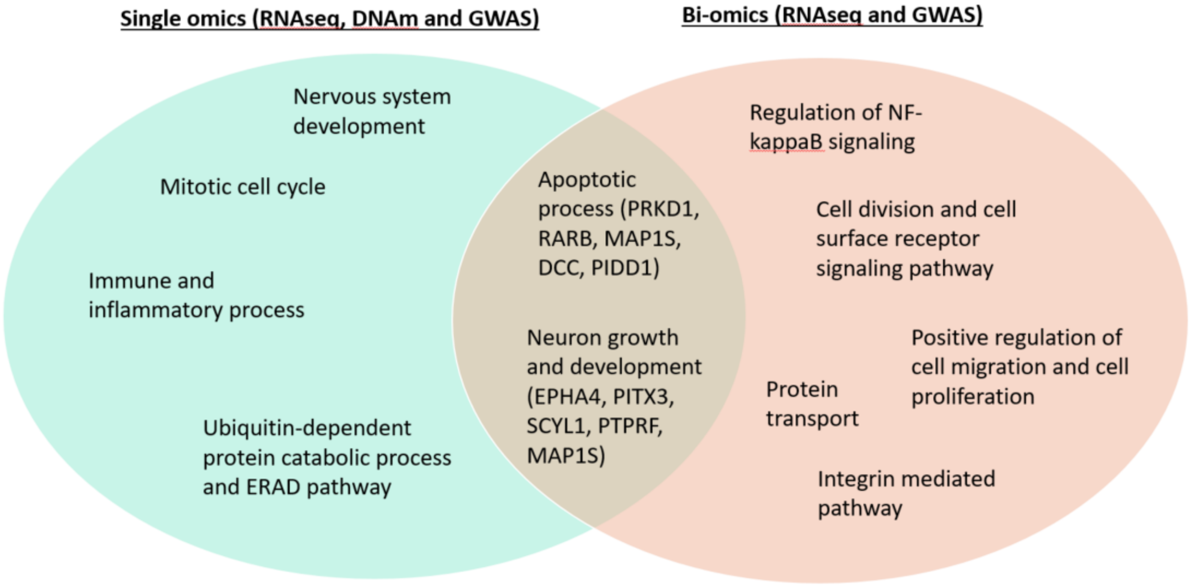

Single omics (RNAseq, DNAm and GWAS)
Bi-omics (RNAseq and GWAS)
Nervous system development
Mitotic cell cycle
Immune and inflammatory process
Ubiquitin-dependent protein catabolic process and ERAD pathway
Apoptotic process (PRKD1, RARB, MAP1S, DCC, PIDD1)
Neuron growth and development (EPHA4, PITX3, SCYL1, PTPRF, MAP1S)
Regulation of NF-kappaB signaling
Cell division and cell surface receptor signaling pathway
Positive regulation of cell migration and cell proliferation
Protein transport
Integrin mediated pathway

Figure 13: Biological process shared by top weighted genes from single-omics and multi-omics E-VAE classifier model

The biological interpretation of latent embedding from single omics E-VAE classifier models found common biological processes in all three single omics E-VAE classifier model such as nervous system development, neuron growth and development, apoptotic process, mitotic cell cycle, immune and inflammatory process and ubiquitin-dependent protein catabolic process and ERAD pathway as shown in figure 13. In the multi-omics model, we observed biological processes common to GWAS SNPs and RNAseq DEGs such as neuron growth and development, apoptotic process, regulation of NF-kappaB signaling, cell division and cell surface receptor signaling pathway, protein transport, integrin mediated pathway and positive regulation of cell migration and cell proliferation. Among the top-weighted features, we found that single omics and multi omics models had a few common GWAS SNPs on the genes ARFGEF2, JMJD1C, PHF20, SFMBT1, TSBP1-AS1 and RNAseq DEGs AC004490.1, RP11-638I2.9, TMA7, TMEM176A as shown in supplementary table 1 and 2 and some distinct genes in single omics and multi omics models. As shown in figure 13, we found two biological processes common between single omics and multi-omics models, apoptotic process and neuron growth and development. We observed several biological process unique to single omics and multi omics models. We found that along with top genes in the amyloid pathophysiology, genes involved in neuronal growth and development were represented in the list of top genes such as, Ephrin type-A receptor 1 (*EPHA1*) and Paired Like Homeodomain 3 (*Pitx3).* At least two previous studies have found an association between genetic variants in EPHA1 and cognitive scores. In a longitudinal study of Caucasians, the most significant nine SNPs at late onset AD (LOAD) risk GWAS loci were genotyped using Taqman assay to examine their association with memory scores. The study found that CLU had a constant effect on memory decline, while ABCA7 and EPHA1 may have a fluctuating effect on memory outcomes [128]. Another study in ADNI cohort identified a SNP on *EPHA1* (rs11771145) has a role in the pathological alteration of the hippocampus and the lateral occipitotemporal and inferior temporal gyri leading to a lower risk of AD [129]. Similarly, recent experimental and mouse models studied in *Pitx3* in association with Parkinson's disease, showed that

Pitx3 gene knockdown decreased *Pitx3* protein expression [130] and *Pitx3* deficiency leads to damages in striatal medium spiny neurons during aging along with nuclear DNA damages [131]. But the role of *Pitx3* gene in association with AD/ADRD has not been reported before and provides opportunities to evaluate a novel gene associated with AD/ADRD.

The network topology analysis performed using top-weighted genes had APP, TRIM25, RNF4, NTRK1 etc. in the list of top-ranked neighbors in the protein-protein interaction (PPI) network. Many variants of APP gene has been associated with early-onset AD [132]. The dysfunction of TRIM25 and RNF4 was associated with amyloidosis and neurodegeneration shows the protective role of these genes in the pathways that underlie AD [133, 134]. The NTRK1 gene variants has been studied in AD pathogenesis specifically for its role in neuron growth and development[135]. There is increasing evidence suggesting the role of neurotrophic factors (NGFB and BDNF) and their receptors (NTRK1, NTRK2, NGFR etc.) in AD pathogenesis [135, 136] but has not been explored in large population-based studies. In an Italian population study of mix of sporadic and familial AD cases and healthy participants. They have genotyped 21 SNPs mapping on candidate genes (4 on NGFB, 5 on NTRK1, 6 on BDNF, 2 on NTRK2 and 4 on NGFR) and showed a strong allele and genotype association of a SNP on NTRK1 with early onset sporadic AD [135]. These study findings are supportive of the novel genes identified in the multi-omics E-VAE classifier model. Thus, the neuronal development pathway with genes such as PITX3, EPHA1 and NTRK1 represents a novel pathway that can be explored further in future studies. Compared to other ML methods, the biological interpretability of E-VAE classifier allowed us to translate the latent representation from integrated multi-omics data can help understand the complex biology in dementia [54] and characterize the disease status.

5.4.2. Strengths and limitations

The HRS surveys were tailored to identify factors contributing to the development and characterization of cognitive impairment and dementia among older adults of 51 years and older. The DL algorithms can handle batch effects and non-linearity observed in high dimensional heterogeneous high throughput data while building a computational model for integrated omics data analysis. The application of E-VAE classifier model to integrate multi-omics data from blood to improve the prediction of dementia is novel. The biological interpretability of latent features generated from multi-omics based deep learning model is another innovative approach. The top weighted features extracted from multi-omics E-VAE classifier model allowed us to translate the deep learning-based features to biological understanding in dementia etiology and generate new hypothesis to be investigated. The model we developed accounted for higher order interactions and noise in multi-omics data in the high dimensional feature space compared to classical dimensionality reduction algorithms. The availability of blood-based data on whole genome-wide transcriptome and genomic variants from large representative population of older adults from two independent studies allowed for better generalizability of our prediction model.

A prediction model trained using omics data from one study and validated in an independent study is hard to implement due to missing data, technical variation in the measurement of omics data, and the definition of outcome variables, which might lead to reduced performance of the ML model. One of the major limitation of our study is that we used only those features from SNPs and DEGs that were common to both the HRS and ROSMAP study. This limited our ability to use a larger number of features available in the HRS to train the multi-omics E-VAE classifier.

In summary, this is the first study developed a framework to integrate multi-omics data from blood to predict AD/ADRD using an end to end deep neural network architecture. The **holistic approach of integrating hierarchical multi-omics data** using a deep neural network model accounted for the higher order interactions within and between various functional omics and allowed us to understand the complex and unknown biology in AD/ADRD [37]. The analysis framework we developed can be easily adapted to integrate other omics data and to study other health outcomes. The knowledge gained from the multi-omics

E-VAE classifier model can be further studied to develop blood-based signatures for the early identification of AD dementia in older adults and will be critical for identifying novel targets for intervention.

Chapter 6. Conclusions and future directions

Clinical diagnosis of AD is often a diagnosis of exclusion and there are no reliable imaging or fluid-based biomarkers that can routinely use in clinical practice to diagnose patients with AD. To realize the goal of personalized medicine in the early identification of AD/ADRD or in AD progression, innovative analysis methods that integrate high-throughput omics data from multiple levels are needed to develop a non-invasive method to provide the objective classification of individuals with and without dementia. The recent availability of blood-based multi-omics data in the HRS cohort, allowed for a holistic evaluation of the heterogeneity and complex nature of AD/ADRD. The HRS surveys were tailored to identify factors contributing to the development and characterization of cognitive impairment and dementia among older adults of 51 years and older. To reduce the challenges that arise from heterogeneity and high dimensionality of multi-omics data we proposed to use an end to end deep learning based integrated multi-omics analysis framework. The integration of various omics data from different levels of regulation and using dimensionality reduction approaches can provide a more comprehensive picture of the downstream changes associated with dementia. Compared to other ML methods, more precise and definitive features identified using a VAE can help understand the complex biology in AD/ADRD [54] and characterize the disease status.

In this dissertation we have demonstrated a framework for integrated multi-omics approach to classify dementia. To achieve this, we developed an explainable variational autoencoder (E-VAE) classifier model that allowed us to analyze each omics separately and also allowed for an integrated omics analysis. We used genetic variants, transcriptome and epigenetic data from blood samples in the Health and Retirement Study (HRS) to develop the single omics and multi-omics E-VAE classifier models. We performed various feature engineering and feature selection approaches on the input omics data to optimize the ML model. As previously reported feature selection strategies and feature engineering are two crucial components of analyzing high-dimensional omics data. There is a saying that goes “A machine learning model is only as good as the data it is fed". We employed a simple yet effective feature selection methods using association analysis with the outcome. We then compared the performance of models using different feature selection methods against those that incorporated all features. Our findings revealed that feature selection is advantageous for learning the E-VAE classifier model. The E-VAE classifier model developed can account for non-linear interactions between and within functional omics. We also observed that hyper parameter tuning can substantially improve the E-VAE classifier prediction performance.

With the explainable model, we were able to identify previously studied and novel multi-omics-based blood biomarkers for the early detection of Alzheimer's disease. Further targeted functional genomics approaches are necessary to confirm the role of genes from our study in association with the pathogenesis of AD/ ADRD. The biological information extracted from the latent embedding of high-dimensional omics can be potentially be used to develop blood-based signatures for early identification of AD/ADRD.

6.1. Innovation

Though VAE models have been applied successfully in other health domains such as cancer, the application of VAE in integrating multi-omics data to improve the prediction of dementia is novel. The biological interpretability of multi-omics latent features generated from a deep learning-based classifier model is also an innovative approach. This was useful to translate the deep learning-based features to novel biological insights in dementia. In addition, the interpretation of the biological meaning of latent features lead to new biological hypothesis to be investigated. Availability of blood-based multi-omics data from a large representative population of older adults allows for better generalizability of study results.

6.2. Summary

Dementia etiology is complex and is influenced by nonlinear and redundant interactions of genes, when a specific pathway alters from the normal these changes are captured in a genome, epigenome and transcriptome. Utilizing a multi-omics E-VAE classifier model optimized to predict dementia, we have extracted a biologically relevant latent representations from high dimensional multi-modal data. The E-

VAE classifier model accounted for the higher order feature interactions and noise in genome-wide data in the high dimensional space compared to classical dimensionality reduction algorithms.

This dissertation developed single omics-based E-VAE classifier models to predict dementia to evaluate the independent explanatory power of each omics model. We expanded the E-VAE classifier model to integrate multi-omics data and evaluated the biological interpretability of latent embedding from multi-omics feature space. This dissertation further our understanding of complex interactions in biological pathways across different levels of regulation and developed a framework that can integrate multi-omics data accounting for complex and non-linear interaction between features at different levels of biological regulation. A major challenge in this dissertation was to deal with imbalanced datasets due to the low prevalence of dementia in the general population. When dealing with imbalanced datasets, where one class is significantly underrepresented compared to the other, accuracy alone can be misleading. This is because a model that always predicts the majority class can still achieve high accuracy, but it will not be useful in practical applications where the minority class is of interest. The AUC-ROC, on the other hand, considers the performance of a model across all possible decision thresholds, making it a more appropriate metric in imbalanced scenarios. It measures the ability of a model to correctly rank positive and negative examples, regardless of the specific classification threshold used. A higher AUC-ROC score indicates better performance in distinguishing between the positive and negative classes. However, it is important to note that while AUC-ROC is useful in evaluating the overall performance of a model in imbalanced datasets, it should not be used as the only criterion for model selection. So we evaluated both sensitivity/ recall and AUC-ROC in order to obtain a more complete picture of the model's performance. As shown in Table 6, we evaluated the performance of independent single omics-based E-VAE classifier models, trained using GWAS SNPs (Chapter 2) and RNAseq DEGs (Chapter 3), as well as a multi-omics integrated E-VAE classifier model using both GWAS SNPs and RNAseq DEGs (Chapter 5). Our analysis revealed that the RNAseq DEG-based E-VAE classifier model exhibited higher sensitivity and AUC in classifying dementia, as compared to the GWAS SNPs-based E-VAE classifier model and the multi-omics model though the differences were not statistically significant. While the RNAseq DEG-based E-VAE classifier model achieved higher sensitivity in classifying dementia in the external validation dataset from the ROSMAP study, the multi-omics based E-VAE classifier model achieved a higher AUC though the differences were not statistically significant.

Table 6: Comparison of prediction performances between the single omics and multi-omics E-VAE classifier models in the HRS and ROSMAP study.

Models	**GWAs SNPs**		**RNAseq**		**Multi-omics**	
	Sensitivity/ Recall (95% CI)	**AUC (95% CI)**	**Sensitivity/ Recall (95% CI)**	**AUC (95% CI)**	**Sensitivity/ Recall (95% CI)**	**AUC (95% CI)**
HRS: E-VAE classifier model (Test (N=52)).	0.73 (0.56, 0.90)	0.69 (0.54, 0.83)	0.77 (0.61, 0.93)	0.82 (0.70, 0.93)	0.73 (0.56, 0.90)	0.80 (0.68, 0.92)
ROSMAP: E-VAE classifier model (Test (N=234))	0.58 (0.30, 0.86)	0.63 (0.46, 0.80)	0.92 (0.76, 1.00)	0.60 (0.43, 0.78)	0.75 (0.51, 0.995)	0.66 (0.49, 0.84)
HRS: Penalized LR model. (Test (N=52))	0.77 (0.61, 0.93)	0.72 (0.58, 0.86)	0.81 (0.66, 0.96)	0.62 (0.46, 0.77)	0.77 (0.61, 0.93)	0.72 (0.58, 0.86)
ROSMAP: Penalized LR model (Test (N=234))	0.17 (0.00, 0.38)	0.43 (0.27, 0.59)	0.17 (0.00, 0.38)	0.35 (0.21, 0.50)	0.33 (0.07, 0.60)	0.56 (0.39, 0.73)

In summary, our study suggests that integrated multi-omics-based E-VAE classifier model had better generalizability and non-random classification performance. These findings need to be confirmed in future studies that are adequately powered to evaluate the differences in sensitivity and AUC-ROC observed using the different approaches.

The knowledge gained from this study generated new hypothesis that can be evaluated in future studies. The integrated multi-omics analysis framework developed can be adapted to study similar outcomes utilizing various omics data. The E-VAE classifier model developed can account for the higher-order interactions in high dimensional omics data. The multi-omics-based deep learning prediction model for dementia improved our understanding of complex interactions in biological pathways across different levels of regulation. This new information can potentially be used to develop blood-based signatures for early identification of AD/ADRD in older adults and will also be critical for identifying novel targets for intervention.

Future directions

The biologically relevant features we identified can be used only to identify the cross-sectional association with dementia. The HRS will have longitudinal measures of cognition available later in 2024. We can then evaluate the performance of the already trained model to predict incident dementia and can estimate the hazard ratios for incident dementia based on the biologically relevant features from multi-omics analysis. We have implemented a balanced class approach to train the E-VAE classifier model that significantly improved the sensitivity of the model to correctly classify dementia cases. But this approach has limited our sample size to train the E-VAE classifier model. We consider to train the E-VAE classifier model using more sample size when more data available in the HRS.

In conclusion, this is the first study integrating whole genome-wide multi-omics data from blood among a large nationally representative cohort of older adults to classify dementia. The deep learning-based model we developed helped us understand the complex biology in AD/ADRD. The knowledge gained from this prediction model can be used to develop blood-based signatures for the early identification of AD/ADRD in older adults and can be critical for identifying novel targets for intervention.

Funding information:

This work was supported by the National Institute on Aging (R01 AG AG060110). The Health and Retirement Study is supported by NIA U01 AG009740.

Conflict of interest statement:

There is no competing financial or non-financial interests in relation to this project.

Availability of data:

Data used in this study can be accessed from the Health and Retirement Study website (https://hrsdata.isr.umich.edu/data-products/public-survey-data?_ga=2.99940638.1654882771.1674664830-1581526203.1653334037). The cognition measures and basic demographic data used in this study is publicly available to registered members of the Health and Retirement Study. The genetic variants data can be accessed by authorized investigators and data are distributed through the NCBI Database of Genotypes and Phenotypes (dbGaP), the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS), and through the HRS web site (https://hrs.isr.umich.edu/data-products/genetic-data). The RNAseq data from the Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) (ROSMAP) obtained from the AD knowledge Portal using the Synapse account (https://www.synapse.org/#!Synapse:syn3191087). We accessed the ROSMAP study cognitive function measures and basic demographics data using a DUA between U of Minnesota and RUSH Alzheimer's Data Center (RADC).

Ethics approval and consent to participate: All study participants in the HRS and ROSMAP study were consented for study participation and data sharing for research use. These studies were approved by the Institutional Review Board at the University of Michigan, Ann Arbor and the Institutional Review Board of Rush University Medical Center respectively.

Supplementary table 1: List of top weighted genes and their weights from respective single omics E-VAE classifier models

GWAS SNPs E-VAE classifier model		**RNAseq DEG E-VAE classifier model**		**DNAm E-VAE classifier model**	
Top weighted genes	**z-score weight**	**Top weighted genes**	**z-score weight**	**Top weighted genes**	**z-score weight**
ARFGEF2	2.84	AC004490.1	2.23	BICD1	2.41
ATXN1L	2.63	BOP1	2.27	ST3GAL3	2.49
BTN1A1	2.74	C12orf73	2.21	HDAC4	2.42
CR1	2.60	CROCC2	2.23	EPHA4	2.46
CSE1L-DT	2.60	CXCR5	2.22	CHMP7	2.42
DPP4	2.65	DLEU7	2.21	SLC9A1	2.41
HMGN4	2.70	MAP1S	2.31	GATS	2.43
JMJD1C	2.65	MICD	2.25	PDZRN4	2.40
LONRF2	2.64	PFKFB2	2.27	RARB	2.46
MTMR4	2.63	RNF175	2.36	PVRIG	2.43
NEK4	2.62	RP11-153M7.3	2.42		
PHF20	2.61	RP11-54A4.2	2.22		
PKD1L3	2.60	RP11-624C23.1	2.31		
PRKD1	2.60	RP11-638I2.9	2.21		
PTPRF	2.69	RP1-93H18.6	2.33		
RBM39	2.62	TMA7	2.22		
RNF43	2.78	TMEM176A	2.21		

SFMBT1	2.60	ZNF865	2.29		
SLC24A4	2.63				
SND1	2.62				
SREBF2	2.69				
TSBP1-AS1	2.61				
TSPOAP1-AS1	2.78				

Supplementary table 2: List of top weighted genes and their weights from multi omics E-VAE classifier model

GWAS SNPs		RNAseq DGE	
Top weighted genes	**z-score weight**	**Top weighted genes**	**z-score weight**
ADAM10	2.30	A2M	2.14
ARFGEF2	2.34	AC004490.1	2.25
CADM2	2.31	AC110781.3	2.13
CPNE1	2.31	CFAP58-AS1	2.18
CSE1L	2.39	CIPC	2.15
DCC	2.36	CSTF3	2.15
HCG11	2.33	CTB-31O20.4	2.12
JMJD1C	2.39	CTD-2575K13.6	2.27
MS4A4A	2.31	DND1P1	2.12
MS4A6E	2.31	DST	2.10
PBRM1	2.32	EPHA1	2.12
PHF20	2.30	FSIP2	2.18
PITX3	2.38	LRR1	2.20
SEPTIN4-AS1	2.34	LY6E	2.25
SFMBT1	2.31	MTX1P1	2.11
ST3GAL3	2.32	MYO5C	2.12
STAU1	2.32	PIDD1	2.13
SYPL2	2.31	RABEP2	2.11
TEX14	2.42	RP11-1100L3.8	2.19

TSBP1-AS1	2.36	RP11-1151B14.4	2.26
		RP11-254F7.2	2.13
		RP11-638I2.9	2.23
		RP11-686G8.1	2.13
		SCYL1	2.10
		SIGLEC10	2.32
		SSC4D	2.14
		TCOF1	2.17
		TMA7	2.15
		TMEM176A	2.11
		TOM1	2.12
		UBE2S	2.19

Supplementary table 3: Biological interpretation of top-weighted features extracted from multi omics E-VAE classifier models

Gene Ontology Process from NCBI DB	GWAS SNPs	RNAseq DEGs
involved_in apoptotic process	DCC	PIDD1
involved_in cell division	TEX14	UBE2S
involved_in cell surface receptor signaling pathway	MS4A6E	EPHA1, LY6E
involved_in integrin-mediated signaling pathway	ADAM10	DST
involved_in negative regulation of DNA-templated transcription	SFMBT3, SFMBT4	CIPC
involved_in neuron projection and development	PITX3	SCYL1
involved_in positive regulation of cell migration	ADAM10	EPHA1
involved_in protein transport	ARFGEF2	RABEP2, TOM1
involved_in regulation of I-kappaB kinase/NF-kappaB signaling	CPNE1	PIDD1
involved_in cell adhesion		DST, SIGLEC10
involved_in chromatin organization	JMJD1C, PHF20, SFMBT2	
involved_in endocytosis		RABEP2, SSC4D, TOM1
involved_in peptidyl-tyrosine phosphorylation		EPHA1, SCYL1
involved_in positive regulation of DNA-templated transcription	JMJD1C, PITX3	
involved_in regulation of transcription by RNA polymerase II	JMJD1C, PBRM1, PHF20, PITX3	

involved_in signal transduction		RABEP2, TOM1, PIDD1

References:

1. Price JL, Morris JC. Tangles and plaques in nondemented aging and “preclinical” Alzheimer's disease. *Annals of Neurology*. 1999;45(3):358-368. doi:https://doi.org/10.1002/1531-8249(199903)45:3<358::AID-ANA12>3.0.CO;2-X
2. Jack CR, Knopman DS, Jagust WJ, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*. 2010/01/01/ 2010;9(1):119-128. doi:https://doi.org/10.1016/S1474-4422(09)70299-6
3. Zierer J, Menni C, Kastenmüller G, Spector TD. Integration of ‘omics’ data in aging research: from biomarkers to systems biology. *Aging Cell*. 2015;14(6):933-944. doi:https://doi.org/10.1111/acel.12386
4. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. May 5 2017;18(1):83. doi:10.1186/s13059-017-1215-1
5. Badhwar A, McFall GP, Sapkota S, et al. A multiomics approach to heterogeneity in Alzheimer's disease: focused review and roadmap. *Brain : a journal of neurology*. May 1 2020;143(5):1315-1331. doi:10.1093/brain/awz384
6. Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*. 2020;doi:10.1093/bib/bbaa177
7. Hess M, Hackenberg M, Binder H. Exploring generative deep learning for omics data using log-linear models. *Bioinformatics*. 2020;doi:10.1093/bioinformatics/btaa623
8. https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet.
9. https://www.alz.org/media/documents/alzheimers-facts-and-figures-2019-r.pdf.
10. Rajan KB, Weuve J, Barnes LL, McAninch EA, Wilson RS, Evans DA. Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020-2060). *Alzheimer's & dementia : the journal of the Alzheimer's Association*. Dec 2021;17(12):1966-1975. doi:10.1002/alz.12362
11. Matthews KA, Xu W, Gaglioti AH, et al. Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015–2060) in adults aged ≥65 years. https://doi.org/10.1016/j.jalz.2018.06.3063. *Alzheimer's & Dementia*. 2019/01/01 2019;15(1):17-24. doi:https://doi.org/10.1016/j.jalz.2018.06.3063
12. Irwin K, Sexton C, Daniel T, Lawlor B, Naci L. Healthy Aging and Dementia: Two Roads Diverging in Midlife? Review. *Frontiers in Aging Neuroscience*. 2018-September-19 2018;10(275)doi:10.3389/fnagi.2018.00275
13. Jack CR, Jr., Bennett DA, Blennow K, et al. A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology*. 2016;87(5):539-547. doi:10.1212/WNL.0000000000002923
14. Tatebe H, Kasai T, Ohmichi T, et al. Quantification of plasma phosphorylated tau to use as a biomarker for brain Alzheimer pathology: pilot case-control studies including patients with Alzheimer’s disease and down syndrome. *Molecular Neurodegeneration*. 2017/09/04 2017;12(1):63. doi:10.1186/s13024-017-0206-8
15. Jiao B, Liu H, Guo L, et al. Performance of Plasma Amyloid β, Total Tau, and Neurofilament Light Chain in the Identification of Probable Alzheimer's Disease in South China. *Front Aging Neurosci*. 2021;13:749649. doi:10.3389/fnagi.2021.749649

16. Shen X-N, Huang S-Y, Cui M, et al. Plasma Glial Fibrillary Acidic Protein in the Alzheimer Disease Continuum: Relationship to Other Biomarkers, Differential Diagnosis, and Prediction of Clinical Progression. *Clinical Chemistry*. 2023;69(4):411-421. doi:10.1093/clinchem/hvad018

17. Stocker H, Beyer L, Perna L, et al. Association of plasma biomarkers, p-tau181, glial fibrillary acidic protein, and neurofilament light, with intermediate and long-term clinical Alzheimer's disease risk: Results from a prospective cohort followed over 17 years. https://doi.org/10.1002/alz.12614. *Alzheimer's & Dementia*. 2023/01/01 2023;19(1):25-35. doi:https://doi.org/10.1002/alz.12614

18. Baiardi S, Quadalti C, Mammana A, et al. Diagnostic value of plasma p-tau181, NfL, and GFAP in a clinical setting cohort of prevalent neurodegenerative dementias. *Alzheimer's research & therapy*. Oct 12 2022;14(1):153. doi:10.1186/s13195-022-01093-6

19. Dubois B, Hampel H, Feldman HH, et al. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimer's & dementia : the journal of the Alzheimer's Association*. 2016;12(3):292-323. doi:10.1016/j.jalz.2016.02.002

20. TOPMed. https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program

21. Greenwood AK, Montgomery KS, Kauer N, et al. The AD Knowledge Portal: A Repository for Multi-Omic Data on Alzheimer's Disease and Aging. *Current Protocols in Human Genetics*. 2020;108(1):e105. doi:https://doi.org/10.1002/cphg.105

22. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious Orders Study and Rush Memory and Aging Project. *Journal of Alzheimer's disease : JAD*. 2018;64(s1):S161-S189. doi:10.3233/JAD-179939

23. De Jager PL, Ma Y, McCabe C, et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci Data*. Aug 7 2018;5:180142. doi:10.1038/sdata.2018.142

24. Bekris LM, Yu CE, Bird TD, Tsuang DW. Genetics of Alzheimer disease. *Journal of geriatric psychiatry and neurology*. Dec 2010;23(4):213-27. doi:10.1177/0891988710383571

25. Pimenova AA, Raj T, Goate AM. Untangling Genetic Risk for Alzheimer's Disease. *Biological psychiatry*. Feb 15 2018;83(4):300-310. doi:10.1016/j.biopsych.2017.05.014

26. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer’s disease risk. *Nature Genetics*. 2019/03/01 2019;51(3):404-413. doi:10.1038/s41588-018-0311-9

27. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics*. 2013;45(12):1452-1458. doi:10.1038/ng.2802

28. Liu C-C, Liu C-C, Kanekiyo T, Xu H, Bu G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat Rev Neurol*. 2013;9(2):106-118. doi:10.1038/nrneurol.2012.263

29. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer’s disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nature Genetics*. 2019/03/01 2019;51(3):414-430. doi:10.1038/s41588-019-0358-2

30. Giri M, Zhang M, Lü Y. Genes associated with Alzheimer's disease: an overview and current status. *Clin Interv Aging*. 2016;11:665-681. doi:10.2147/CIA.S105769

31. Ridge PG, Hoyt KB, Boehme K, et al. Assessment of the genetic variance of late-onset Alzheimer's disease. *Neurobiology of Aging*. 2016/05/01/ 2016;41:200.e13-200.e20. doi:https://doi.org/10.1016/j.neurobiolaging.2016.02.024

32. Liu J, Zhao W, Ware EB, Turner ST, Mosley TH, Smith JA. DNA methylation in the APOE genomic region is associated with cognitive function in African Americans. *BMC Medical Genomics*. 2018/05/08 2018;11(1):43. doi:10.1186/s12920-018-0363-9

33. De Jager PL, Srivastava G, Lunnon K, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nature neuroscience*. Sep 2014;17(9):1156-63. doi:10.1038/nn.3786

34. Darst BF, Lu Q, Johnson SC, Engelman CD. Integrated analysis of genomics, longitudinal metabolomics, and Alzheimer's risk factors among 1,111 cohort participants. *bioRxiv*. 2018:436923. doi:10.1101/436923

35. HRS. https://hrsdata.isr.umich.edu/data-products/sensitive-health?_ga=2.232134431.1090083975.1629121812-411200660.1601442014

36. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and biology insights*. 2020;14:1177932219899051. doi:10.1177/1177932219899051

37. Nguyen ND, Wang D. Multiview learning for understanding functional multiomics. *PLoS Comput Biol*. Apr 2020;16(4):e1007677. doi:10.1371/journal.pcbi.1007677

38. Clark C, Dayon L, Masoodi M, Bowman G, Popp J. An Integrative Multi-omics Approach Reveals New Central Nervous System Pathway Alterations in Alzheimer's Disease. Research Square; 2020.

39. Xicota L, Ichou F, Lejeune F-X, et al. Multi-omics signature of brain amyloid deposition in asymptomatic individuals at-risk for Alzheimer's disease: The INSIGHT-preAD study. *EBioMedicine*. 2019/09/01/ 2019;47:518-528. doi:https://doi.org/10.1016/j.ebiom.2019.08.051

40. Pinu FR, Beale DJ, Paten AM, et al. Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. *Metabolites*. 2019;9(4):76. doi:10.3390/metabo9040076

41. Pimplikar SW. Multi-omics and Alzheimer's disease: a slower but surer path to an efficacious therapy? *American journal of physiology Cell physiology*. Jul 1 2017;313(1):C1-c2. doi:10.1152/ajpcell.00109.2017

42. Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. *Front Oncol*. 2020;10:1030. doi:10.3389/fonc.2020.01030

43. Patel-Murray NL, Adam M, Huynh N, Wassie BT, Milani P, Fraenkel E. A Multi-Omics Interpretable Machine Learning Model Reveals Modes of Action of Small Molecules. *Scientific Reports*. 2020/01/22 2020;10(1):954. doi:10.1038/s41598-020-57691-7

44. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. May 28 2015;521(7553):436-44. doi:10.1038/nature14539

45. Koteluk O, Wartecki A, Mazurek S, Kołodziejczak I, Mackiewicz A. How Do Machines Learn? Artificial Intelligence as a New Era in Medicine. *J Pers Med*. 2021;11(1):32. doi:10.3390/jpm11010032

46. Lek S, Park YS. Artificial Neural Networks. In: Jørgensen SE, Fath BD, eds. *Encyclopedia of Ecology*. Academic Press; 2008:237-245.

47. Kingma DP, Welling M. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*. 2019;12(4):307-392. doi:10.1561/2200000056

48. Kingma DP, Welling M. Auto-Encoding Variational Bayes. *CoRR*. 2014;abs/1312.6114

49. Wei R, Garcia C, El-Sayed A, Peterson V, Mahmood A. Variations in Variational Autoencoders - A Comparative Evaluation. *IEEE Access*. 2020;8:153651-153670.

50. Chan CK, Hadjitheodorou A, Tsai TYC, Theriot JA. Quantitative comparison of principal component analysis and unsupervised deep learning using variational autoencoders for shape analysis of motile cells. *bioRxiv*. 2020:2020.06.26.174474. doi:10.1101/2020.06.26.174474

51. <An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos.pdf>.

52. Machiraju G. Multi-omics factorization illustrates the added value of deep learning approaches. 2019:

53. Ma T, Zhang A. Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). *BMC Genomics*. 2019/12/20 2019;20(11):944. doi:10.1186/s12864-019-6285-x

54. Zhang L, Lv C, Jin Y, et al. Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. Original Research. *Frontiers in Genetics*. 2018-October-18 2018;9(477)doi:10.3389/fgene.2018.00477

55. Zhang X, Zhang J, Sun K, Yang X, Dai C, Guo Y. Integrated Multi-omics Analysis Using Variational Autoencoders: Application to Pan-cancer Classification. 2019:765-769.

56. Svensson V, Gayoso A, Yosef N, Pachter L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*. Jun 1 2020;36(11):3418-3421. doi:10.1093/bioinformatics/btaa169

57. Seninge L, Anastopoulos I, Ding H, Stuart J. Biological network-inspired interpretable variational autoencoder. *bioRxiv*. 2020:2020.12.17.423310. doi:10.1101/2020.12.17.423310

58. Way GP, Greene CS. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac Symp Biocomput*. 2018;23:80-91.

59. Lee T, Lee H. Prediction of Alzheimer's disease using blood gene expression data. *Sci Rep*. Feb 26 2020;10(1):3485. doi:10.1038/s41598-020-60595-1

60. Zheng-Lin T, Hai-Xing W, Shao-Xun Y, Xiao S, Jian-Ming X. Classification of Alzheimer's Disease Based on Stacked Denoising Autoencoder. 2018:248-253.

61. Ren J, Zhang B, Wei D, Zhang Z. Identification of Methylated Gene Biomarkers in Patients with Alzheimer's Disease Based on Machine Learning. *Biomed Res Int*. 2020;2020:8348147-8348147. doi:10.1155/2020/8348147

62. Jo T, Nho K, Risacher SL, Saykin AJ, for the Alzheimer's Neuroimaging I. Deep learning detection of informative features in tau PET for Alzheimer's disease classification. *BMC Bioinformatics*. 2020/12/28 2020;21(21):496. doi:10.1186/s12859-020-03848-0

63. Klein H-U, Schäfer M, Bennett DA, Schwender H, De Jager PL. Bayesian integrative analysis of epigenomic and transcriptomic data identifies Alzheimer's disease candidate genes and networks. *PLoS computational biology*. 2020;16(4):e1007771-e1007771. doi:10.1371/journal.pcbi.1007771

64. Tasaki S, Gaiteri C, Mostafavi S, et al. Multi-omic Directed Networks Describe Features of Gene Regulation in Aged Brains and Expand the Set of Genes Driving Cognitive Decline. *Front Genet*. 2018;9:294. doi:10.3389/fgene.2018.00294

65. Bellenguez C, Küçükali F, Jansen IE, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet*. 2022/04/01 2022;54(4):412-436. doi:10.1038/s41588-022-01024-z

66. Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *The American Journal of Human Genetics*. 2001/07/01/ 2001;69(1):138-147. doi:https://doi.org/10.1086/321276

67. Heath L, Earls JC, Magis AT, et al. Manifestations of Alzheimer's disease genetic risk in the blood are evident in a multiomic analysis in healthy adults aged 18 to 90. *Scientific Reports*. 2022/04/12 2022;12(1):6117. doi:10.1038/s41598-022-09825-2

68. Gatz M, Reynolds CA, Fratiglioni L, et al. Role of genes and environments for explaining Alzheimer disease. *Archives of general psychiatry*. Feb 2006;63(2):168-74. doi:10.1001/archpsyc.63.2.168

69. Romero-Rosales BL, Tamez-Pena JG, Nicolini H, Moreno-Treviño MG, Trevino V. Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modeling. *PloS one*. 2020;15(4):e0232103. doi:10.1371/journal.pone.0232103

70. Monk B, Rajkovic A, Petrus S, Rajkovic A, Gaasterland T, Malinow R. A Machine Learning Method to Identify Genetic Variants Potentially Associated With Alzheimer's Disease. Original Research. *Frontiers in Genetics*. 2021-June-14 2021;12doi:10.3389/fgene.2021.647436

71. You J, Zhang YR, Wang HF, et al. Development of a novel dementia risk prediction model in the general population: A large, longitudinal, population-based machine-learning study. *EClinicalMedicine*. Nov 2022;53:101665. doi:10.1016/j.eclinm.2022.101665

72. Cruchaga C, Del-Aguila JL, Saef B, et al. Polygenic risk score of sporadic late-onset Alzheimer's disease reveals a shared architecture with the familial and early-onset forms. https://doi.org/10.1016/j.jalz.2017.08.013. *Alzheimer's & Dementia*. 2018/02/01 2018;14(2):205-214. doi:https://doi.org/10.1016/j.jalz.2017.08.013

73. Jo T, Nho K, Bice P, Saykin AJ, For The Alzheimer's Disease Neuroimaging I. Deep learning-based identification of genetic variants: application to Alzheimer's disease classification. *Briefings in Bioinformatics*. 2022;23(2):bbac022. doi:10.1093/bib/bbac022

74. Kira K, Rendell LA. A Practical Approach to Feature Selection. In: Sleeman D, Edwards P, eds. *Machine Learning Proceedings 1992*. Morgan Kaufmann; 1992:249-256.

75. Zou KH, O'Malley AJ, Mauri L. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation*. 2007/02/06 2007;115(5):654-657. doi:10.1161/CIRCULATIONAHA.105.594929

76. Crimmins EM, Kim JK, Langa KM, Weir DR. Assessment of cognition using surveys and neuropsychological assessment: the Health and Retirement Study and the Aging, Demographics, and Memory Study. *J Gerontol B Psychol Sci Soc Sci*. 2011;66 Suppl 1(Suppl 1):i162-i171. doi:10.1093/geronb/gbr048

77. Quality Control Report for Genotypic Data. https://hrs.isr.umich.edu/sites/default/files/genetic/HRS-QC-Report-Phase-4_Nov2021_FINAL.pdf.

78. Moore JH. Chapter 5 - Detecting, Characterizing, and Interpreting Nonlinear Gene–Gene Interactions Using Multifactor Dimensionality Reduction. In: Dunlap JC, Moore JH, eds. *Advances in Genetics*. Academic Press; 2010:101-116.

79. Davies G, Lam M, Harris SE, et al. Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat Commun*. May 29 2018;9(1):2098. doi:10.1038/s41467-018-04362-x

80. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet*. Mar 2019;51(3):414-430. doi:10.1038/s41588-019-0358-2

81. Wightman DP, Jansen IE, Savage JE, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat Genet*. Sep 2021;53(9):1276-1282. doi:10.1038/s41588-021-00921-z

82. Hofmann-Apitius M, Ball G, Gebel S, et al. Bioinformatics Mining and Modeling Methods for the Identification of Disease Mechanisms in Neurodegenerative Disorders. *International journal of molecular sciences*. Dec 7 2015;16(12):29179-206. doi:10.3390/ijms161226148

83. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Review. *Frontiers in Bioinformatics*. 2022-June-27 2022;2doi:10.3389/fbinf.2022.927312

84. Sancesario GM, Bernardini S. Alzheimer's disease in the omics era. *Clinical biochemistry*. Sep 2018;59:9-16. doi:10.1016/j.clinbiochem.2018.06.011

85. Ammann JU, Cooke A, Trowsdale J. Butyrophilin Btn2a2 inhibits TCR activation and phosphatidylinositol 3-kinase/Akt pathway signaling and induces Foxp3 expression in T lymphocytes. *Journal of immunology (Baltimore, Md : 1950)*. May 15 2013;190(10):5030-6. doi:10.4049/jimmunol.1203325

86. Redwan EM, Al-Hejin AM, Almehdar HA, Elsaway AM, Uversky VN. Prediction of Disordered Regions and Their Roles in the Anti-Pathogenic and Immunomodulatory Functions of Butyrophilins. *Molecules (Basel, Switzerland)*. Feb 4 2018;23(2)doi:10.3390/molecules23020328

87. Ashton NJ, Pascoal TA, Karikari TK, et al. Plasma p-tau231: a new biomarker for incipient Alzheimer's disease pathology. *Acta neuropathologica*. May 2021;141(5):709-724. doi:10.1007/s00401-021-02275-6

88. Chatterjee P, Pedrini S, Doecke JD, et al. Plasma Aβ42/40 ratio, p-tau181, GFAP, and NfL across the Alzheimer's disease continuum: A cross-sectional and longitudinal study in the AIBL cohort. https://doi.org/10.1002/alz.12724. *Alzheimer's & Dementia*. 2022/07/21 2022;n/a(n/a)doi:https://doi.org/10.1002/alz.12724

89. Chatterjee P, Vermunt L, Gordon BA, et al. Plasma glial fibrillary acidic protein in autosomal dominant Alzheimer's disease: Associations with Aβ-PET, neurodegeneration, and cognition. https://doi.org/10.1002/alz.12879. *Alzheimer's & Dementia*. 2022/12/28 2022;n/a(n/a)doi:https://doi.org/10.1002/alz.12879

90. Beebe-Wang N, Celik S, Weinberger E, et al. Unified AI framework to uncover deep interrelationships between gene expression and Alzheimer's disease neuropathologies. *Nat Commun*. Sep 10 2021;12(1):5369. doi:10.1038/s41467-021-25680-7

91. Idda ML, Munk R, Abdelmohsen K, Gorospe M. Noncoding RNAs in Alzheimer's disease. *Wiley interdisciplinary reviews RNA*. Mar 2018;9(2)doi:10.1002/wrna.1463

92. Li J, Liu C. Coding or Noncoding, the Converging Concepts of RNAs. *Front Genet*. 2019;10:496. doi:10.3389/fgene.2019.00496

93. Bottero V, Potashkin JA. Meta-Analysis of Gene Expression Changes in the Blood of Patients with Mild Cognitive Impairment and Alzheimer’s Disease Dementia. *International journal of molecular sciences*. 2019;20(21). doi:10.3390/ijms20215403

94. Lunnon K, Ibrahim Z, Proitsi P, et al. Mitochondrial Dysfunction and Immune Activation are Detectable in Early Alzheimer's Disease Blood. *Journal of Alzheimer's Disease*. 2012;30:685-710. doi:10.3233/JAD-2012-111592

95. Kursa MB, Jankowski A, Rudnicki WR. Boruta – A System for Feature Selection. *Fundamenta Informaticae*. 2010;101:271-285. doi:10.3233/FI-2010-288

96. Abdullah MN, Wah YB, Abdul Majeed AB, Zakaria Y, Shaadan N. Identification of blood-based transcriptomics biomarkers for Alzheimer's disease using statistical and machine learning classifier. *Informatics in Medicine Unlocked*. 2022/01/01/ 2022;33:101083. doi:https://doi.org/10.1016/j.imu.2022.101083

97. Chiricosta L, D'Angiolini S, Gugliandolo A, Mazzon E. Artificial Intelligence Predictor for Alzheimer's Disease Trained on Blood Transcriptome: The Role of Oxidative Stress. *International journal of molecular sciences*. May 7 2022;23(9)doi:10.3390/ijms23095237

98. Shigemizu D, Mori T, Akiyama S, et al. Identification of potential blood biomarkers for early diagnosis of Alzheimer's disease through RNA sequencing analysis. *Alzheimer's research & therapy*. Jul 16 2020;12(1):87. doi:10.1186/s13195-020-00654-x

99. Wang Q, Chen K, Su Y, Reiman EM, Dudley JT, Readhead B. Deep learning-based brain transcriptomic signatures associated with the neuropathological and clinical severity of Alzheimer's disease. *Brain Communications*. 2022;4(1):fcab293. doi:10.1093/braincomms/fcab293

100. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. Jan 1 2010;26(1):139-40. doi:10.1093/bioinformatics/btp616

101. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1995/01/01 1995;57(1):289-300. doi:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

102. Chen H, He Y, Ji J, Shi Y. A Machine Learning Method for Identifying Critical Interactions Between Gene Pairs in Alzheimer's Disease Prediction. *Frontiers in neurology*. 2019;10:1162. doi:10.3389/fneur.2019.01162

103. Ji J, He D, Feng Y, He Y, Xue F, Xie L. JDINAC: joint density-based non-parametric differential interaction network analysis and classification using high-dimensional sparse omics data. *Bioinformatics*. Oct 1 2017;33(19):3080-3087. doi:10.1093/bioinformatics/btx360

104. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One*. 2014;9(1):e78644. doi:10.1371/journal.pone.0078644

105. Rao MS, Van Vleet TR, Ciurlionis R, et al. Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. *Front Genet*. 2018;9:636. doi:10.3389/fgene.2018.00636

106. Smith AM, Walsh JR, Long J, et al. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics*. 2020/03/20 2020;21(1):119. doi:10.1186/s12859-020-3427-8

107. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*. 2013/01/01 2013;38(1):23-38. doi:10.1038/npp.2012.112

108. Gatz M, Pedersen NL, Berg S, et al. Heritability for Alzheimer's disease: the study of dementia in Swedish twins. *The journals of gerontology Series A, Biological sciences and medical sciences*. Mar 1997;52(2):M117-25. doi:10.1093/gerona/52a.2.m117

109. Trerotola M, Relli V, Simeone P, Alberti S. Epigenetic inheritance and the missing heritability. *Human genomics*. Jul 28 2015;9(1):17. doi:10.1186/s40246-015-0041-3

110. Acha B, Corroza J, de Gordoa JS-R, et al. A blood-based panel of DNA methylation markers improves diagnosis accuracy of Alzheimer's disease. Research Square; 2022.

111. Tannorella P, Stoccoro A, Tognoni G, et al. Methylation analysis of multiple genes in blood DNA of Alzheimer's disease and healthy individuals. *Neuroscience Letters*. 2015/07/23/ 2015;600:143-147. doi:https://doi.org/10.1016/j.neulet.2015.06.009

112. Sugden K, Caspi A, Elliott ML, et al. Association of Pace of Aging Measured by Blood-Based DNA Methylation With Age-Related Cognitive Impairment and Dementia. *Neurology*. 2022;99(13):e1402. doi:10.1212/WNL.0000000000200898

113. Faul JD, Kim JK, Levine ME, Thyagarajan B, Weir DR, Crimmins EM. Epigenetic-based age acceleration in a representative sample of older Americans: Associations with aging-related morbidity and mortality. *Proceedings of the National Academy of Sciences*. 2023/02/28 2023;120(9):e2215840120. doi:10.1073/pnas.2215840120

114. Reed RG, Carroll JE, Marsland AL, Manuck SB. DNA methylation-based measures of biological aging and cognitive decline over 16-years: preliminary longitudinal findings in midlife. *Aging*. Nov 11 2022;14(23):9423-9444. doi:10.18632/aging.204376

115. McCartney DL, Stevenson AJ, Walker RM, et al. Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*. 2018/01/01/ 2018;10:429-437. doi:https://doi.org/10.1016/j.dadm.2018.05.006

116. Sibbett RA, Altschul DM, Marioni RE, Deary IJ, Starr JM, Russ TC. DNA methylation-based measures of accelerated biological ageing and the risk of dementia in the oldest-old: a study of the Lothian Birth Cohort 1921. *BMC psychiatry*. Feb 28 2020;20(1):91. doi:10.1186/s12888-020-2469-9

117. Fransquet PD, Lacaze P, Saffery R, McNeil J, Woods R, Ryan J. Blood DNA methylation as a potential biomarker of dementia: A systematic review. *Alzheimer's & dementia : the journal of the Alzheimer's Association*. Jan 2018;14(1):81-103. doi:10.1016/j.jalz.2017.10.002

118. <HRS Epigenetic Clocks.pdf>.

119. Tian Y, Morris TJ, Webster AP, et al. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics*. 2017;33(24):3982-3984. doi:10.1093/bioinformatics/btx513

120. Chen L. Multi-task deep autoencoder to predict Alzheimer’s disease progression using temporal DNA methylation data in peripheral blood. *medRxiv*. 2022:2022.04.02.22273346. doi:10.1101/2022.04.02.22273346

121. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet*. Jan 2007;39(1):17-23. doi:10.1038/ng1934

122. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer’s disease stage. *Scientific Reports*. 2021/02/05 2021;11(1):3254. doi:10.1038/s41598-020-74399-w

123. Park C, Ha J, Park S. Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. *Expert Systems with Applications*. 2020/02/01/ 2020;140:112873. doi:https://doi.org/10.1016/j.eswa.2019.112873

124. Clark C, Dayon L, Masoodi M, Bowman GL, Popp J. An integrative multi-omics approach reveals new central nervous system pathway alterations in Alzheimer's disease. *Alzheimer's research & therapy*. Apr 1 2021;13(1):71. doi:10.1186/s13195-021-00814-7

125. François M, Karpe AV, Liu J-W, et al. Multi-Omics, an Integrated Approach to Identify Novel Blood Biomarkers of Alzheimer&rsquo;s Disease. *Metabolites*. 2022;12(10). doi:10.3390/metabo12100949

126. Shi L, Xu J, Green R, et al. Multiomics profiling of human plasma and cerebrospinal fluid reveals ATN-derived networks and highlights causal links in Alzheimer's disease. https://doi.org/10.1002/alz.12961. *Alzheimer's & Dementia*. 2023/02/15 2023;n/a(n/a)doi:https://doi.org/10.1002/alz.12961

127. Karikari TK, Pascoal TA, Ashton NJ, et al. Blood phosphorylated tau 181 as a biomarker for Alzheimer's disease: a diagnostic performance and prediction modelling study using data from four prospective cohorts. *The Lancet Neurology*. May 2020;19(5):422-433. doi:10.1016/s1474-4422(20)30071-5

128. Carrasquillo MM, Crook JE, Pedraza O, et al. Late-onset Alzheimer's risk variants in memory decline, incident mild cognitive impairment, and Alzheimer's disease. *Neurobiol Aging*. Jan 2015;36(1):60-7. doi:10.1016/j.neurobiolaging.2014.07.042

129. Wang H-F, Tan L, Hao X-K, et al. Effect of EPHA1 Genetic Variation on Cerebrospinal Fluid and Neuroimaging Biomarkers in Healthy, Mild Cognitive Impairment and Alzheimer's Disease Cohorts. *Journal of Alzheimer's Disease*. 2015;44:115-123. doi:10.3233/JAD-141488
130. Chen J, Kang XY, Tang CX, Gao DS. Impact of Pitx3 gene knockdown on glial cell line-derived neurotrophic factor transcriptional activity in dopaminergic neurons. *Neural regeneration research*. Aug 2017;12(8):1347-1351. doi:10.4103/1673-5374.213557
131. Chen X, Yang Z, Shao Y, et al. Pitx3 deficiency promotes age-dependent alterations in striatal medium spiny neurons. Original Research. *Frontiers in Aging Neuroscience*. 2022-September-07 2022;14doi:10.3389/fnagi.2022.960479
132. Tcw J, Goate AM. Genetics of β-Amyloid Precursor Protein in Alzheimer's Disease. *Cold Spring Harbor perspectives in medicine*. Jun 1 2017;7(6)doi:10.1101/cshperspect.a024539
133. Zhu Y, Afolabi LO, Wan X, Shim JS, Chen L. TRIM family proteins: roles in proteostasis and neurodegenerative diseases. *Open biology*. Aug 2022;12(8):220098. doi:10.1098/rsob.220098
134. Gómez-Tortosa E, Baradaran-Heravi Y, Dillen L, et al. TRIM25 mutation (p.C168*), coding for an E3 ubiquitin ligase, is a cause of early-onset autosomal dominant dementia with amyloid load and parkinsonism. *Alzheimer's & dementia : the journal of the Alzheimer's Association*. Dec 28 2022;doi:10.1002/alz.12913
135. Cozza A, Melissari E, Iacopetti P, et al. SNPs in Neurotrophin System Genes and Alzheimer's Disease in an Italian Population. *Journal of Alzheimer's Disease*. 2008;15:61-70. doi:10.3233/JAD-2008-15105
136. Bathina S, Das UN. Brain-derived neurotrophic factor and its clinical implications. *Archives of medical science : AMS*. Dec 10 2015;11(6):1164-78. doi:10.5114/aoms.2015.56342