Causal Feature Extraction and Inference for High-Dimensional Data

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Yuchen Yao

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Xiaotong Shen and
Wei Pan, Advisers

November 2024

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my advisor, Dr. Xiaotong Shen, and my co-advisor, Dr. Wei Pan, for their invaluable support, guidance, and encouragement during my PhD years. Also, I am profoundly grateful to my thesis committee members, Dr. Adam J. Rothman and Dr. Jie Ding, for their commitment of time reviewing my thesis. Finally, I want to thank my friends, as well as the staff and students at the School of Statistics, for their support and assistance.

Dedication

To my parents, Xiaofeng Yao and Xiaohong Chen.

Abstract

This dissertation investigates the problem of extracting and inferring causal features from high-dimensional data. A series of approaches based on instrumental variable (IV) regression and deep learning techniques are presented in this dissertation to solve this problem.

In the first project, we propose a novel method, deep feature extraction via instrumental variable regression (DeepFEIVR), which utilizes a non-linear neural network to extract causal features from high-dimensional data, such as neuroimaging, to predict outcomes like Alzheimer's disease (AD) status. This new method also preserves a linear relationship between the extracted features and IVs, such as genetic variants. DeepFEIVR not only handles high-dimensional individual-level data for model building but is also compatible with genome-wide association study (GWAS) summary data, enabling the testing of associations between the extracted features and the outcome in subsequent analyses. In addition, we propose an extension, DeepFEIVR-CA, for covariate adjustment (CA). We apply DeepFEIVR and DeepFEIVR-CA to individual-level data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) for model building, followed by their applications to AD GWAS summary data from the International Genomics of Alzheimer's Project (IGAP) and neuroimaging data from the UK Biobank, showcasing how the extracted causal features relate to AD and various brain endophenotypes.

In the second project, we propose a variant of DeepFEIVR with residual inclusion, called DeepFEIVR-RI, and extend both DeepFEIVR and DeepFEIVR-RI to accommodate a large number of related IVs. We apply DeepFEIVR and DeepFEIVR-RI to a large dataset, the UK Biobank, to show that the extracted features (genetic compo-

nents) in electrocardiogram (ECG) recordings are significantly associated with atrial fibrillation (AF) (p-values < 1e-8). Furthermore, we adapt a recent algorithm called dnn-loc, enabling a visual interpretation of connections between specific ECG components and AF through extracted causal features, thus advancing the understanding of AF etiology.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, we begin by introducing the necessary notations and settings used in both projects. After that, we present two classical instrumental variable regression approaches, two-stage least squares (2SLS) and two-stage residual inclusion (2SRI), which serve as the foundation of our proposed methods. Lastly, we provide an overview of this dissertation.

1.1 Notation

Assume that Z represents p IVs , X represents k exposures, and Y represents an outcome, respectively. Both X and Y are influenced by correlated confounders (U_1, U_2). In this setting, we define a training set $\mathcal{D}_{tr} = \{Z_{tr}, X_{tr}, Y_{tr}\}$ with a training size of n_{tr} and a validation set $\mathcal{D}_{val} = \{Z_{val}, X_{val}, Y_{val}\}$ with a validation size of n_{val}. For hypothesis testing, we begin with an individual-level test set $\mathcal{D}_{te}^{indv} = \{Z_{te}, Y_{te}\}$ with a test size of n_{te}. Based on this individual-level test set $\mathcal{D}_{te}^{indv}$, we can then create summary statistics $\mathcal{D}_{te}^{s} = \{(\hat{\gamma}_j, \widehat{\text{Var}}(\hat{\gamma}_j)) : j = 1, ..., p\}$, in which $\hat{\gamma}_j$ is the estimated effect size of the j-th component of Z on Y. It can be estimated by a linear model regressing Y_{te} on the j-th column of Z_{te}. The summary statistics also include $\widehat{\text{Var}}(\hat{\gamma}_j)$ for the j-th IV, which is the squared standard error (variance estimate) of γ_j. Un-

der certain cases, the summary statistics $\mathcal{D}_{te}^{s}$ can be used in hypothesis testing as a substitute for the individual-level test set.

1.2 Two-Stage Least Squares

We begin with introducing 2SLS (Klungel et al. (2015)). Its causal model structure is as follows:

$$\begin{aligned} \text{Stage 1}: \quad & X = ZB + U_1 + \epsilon_1, \\ \text{Stage 2}: \quad & Y = X\beta + U_2 + \epsilon_2, \end{aligned} \tag{1.1}$$

where U_1 and U_2 are two confounders, ϵ_1 and ϵ_2 are two random noises with zero means and constant variances. For notational simplicity, we omit the intercepts in the notation and assume Z, X, Y are already centered at sample mean zero. The parameters to be estimated are $B \in \mathbb{R}^{p\times k}$ and $\beta \in \mathbb{R}^{k}$. Due to the presence of confounders affecting both X and Y, U_1 and U_2 are correlated, while ϵ_1 and ϵ_2 remain independent. In order to eliminate the influence from confounders, Z, as IVs, are required to satisfy the following three assumptions:

1. The distribution of the exposure X, given IVs Z, is not constant in Z.
2. Z is independent of the outcome Y, conditional on X, U_2 and ϵ_2;
3. Z and $(U_1, U_2, \epsilon_1, \epsilon_2)$ are independent.

In Stage 1 of 2SLS, we estimate the mean of X by fitting a linear regression model of X on Z. In Stage 2, we estimate Y by regressing Y on $\hat{X}$, where $\hat{X}$ is the estimated mean of X obtained from Stage 1. In hypothesis testing, the null hypothesis states $\beta_j = 0$ for some or all of $j = 1, 2, ..., k$, where β_j is a component of $\beta = (\beta_1, \beta_2 ..., \beta_k)$. Given that Y and Z are assumed to follow a linear relationship in the 2SLS framework,

we can use summary statistics in hypothesis testing. However, 2SLS is inefficient in modeling high-dimensional exposures by using a linear model in Stage 1. Therefore, in Chapter 2, we consider a deep learning-based instrumental variable regression method, deep feature extraction via instrumental variable regression (DeepFEIVR), to address this problem.

1.3 Two-Stage Residual Inclusion

The model structure of 2SRI (Terza et al. (2008)) assumes

$$\begin{aligned} \text{Stage 1}: &\quad X = ZB + U, \\ \text{Stage 2}: &\quad Y = X\beta + U\alpha + \epsilon. \end{aligned}$$

Based on the model structure and assumptions in 2SLS, we additionally assume that the noise ϵ_1 in Equation (1.1) can be neglected or absorbed into U_1 (U in 2SRI) and U_2 can be written as a linear combination of U_1 (U in 2SRI). Then the working model in Stage 2 is $Y = X\beta + (X - Z\hat{B})\alpha + \epsilon$, where $\hat{B}$ is the estimate of B in Stage 1. As the error term in Stage 2 decreases from $U\alpha + \epsilon$ to ϵ, the estimation variance of β may also be reduced. In Chapter 3, we propose a variant of DeepFEIVR, named DeepFEIVR-RI, inspired by the idea in 2SRI.

1.4 Overview of Dissertation

In causal inference, instrumental variable (IV) regression is widely applied to identify a causal relationship between an exposure and an outcome. However, classical IV regression approaches, such as 2SLS and 2SRI, are inefficient when the exposure is high-dimensional, such as neuroimaging and ECG recordings. A neural network

can serve as a non-linear function to extract features from high-dimensional data efficiently; however, there is no guarantee of causality between these extracted features and the outcome.

To achieve two objectives together, in Chapter 2, we propose DeepFEIVR, a non-linear extension of the classical IV regression model 2SLS, to extract causal features from high-dimensional data and infer their causality with the outcome. Additionally, it preserves a linear relationship between IVs and the outcome to enable hypothesis testing solely utilizing summary statistics. Simulation studies and real data application to the ADNI and IGAP datasets are conducted to demonstrate the inference performance of DeepFEIVR and its ability to handle summary statistics. Subsequently, we provide interpretations of the extracted causal features by using techniques such as gradient-weighted class activation map (Grad-CAM) (Selvaraju et al. (2019)), to demonstrate how the extracted features are related to specific brain regions.

In Chapter 3, DeepFEIVR-residual inclusion (DeepFEIVR-RI) is proposed, aiming to reduce the estimation variance in DeepFEIVR. The efficiency of the extracted causal features greatly relies on the choice of IVs. For genetic data, we propose using block-based polygenic risk scores (PRSs) as IVs to combine information from a great number of single nucleotide polymorphisms (SNPs) and avoid an excessive IV dimension. We compare the performance from DeepFEIVR and DeepFEIVR-RI with various IV choices on the UK Biobank, and demonstrate that block-based PRSs as IVs are more effective than using individual SNPs as IVs. We also adapt a visualization algorithm (dnn-locate) to allow for interpreting causal (genetically imputed) features.

Chapter 2

Deep Causal Feature Extraction and Inference with Neuroimaging Genetic Data

2.1 Introduction

Alzheimer's disease (AD) results in memory loss, dementia, and behavioral changes, causing over 121,000 deaths in 2019 (Apostolova (2016); Association (2022)). In 2022, it was estimated that there were 6.5 million individuals affected by AD among the American elderly aged 65 or above, a number predicted to reach 13.8 million by 2060 (Association (2022)). Due to the prevalence and severe impact of AD, a growing number of researchers focus on AD data collection and analysis. The Alzheimer's Disease Neuroimaging Initiative (ADNI) (Beltran et al. (2020)) dataset is one of the most comprehensive and widely used resources for AD study. A major part of this dataset is structural magnetic resonance imaging (MRI) scans, which are three-dimensional images revealing brain structures and regional conditions. In addition to MRI scans, the ADNI dataset also includes genome-wide SNPs for each individual, which are genetic variants nowadays commonly utilized in genome-wide association studies (GWASs).

Under the reasonable assumption that the pathway of AD progresses from genetics (and environmental factors) to brain atrophy, and then ultimately to AD, it has been advocated to utilize neuroimaging features as endophenotypes. These features, because of their closer proxy to causal genetic factors in the AD pathway, can gain statistical power and motivate the advancements in the field of neuroimaging genetics. In addition to increasing power, the use of endophenotypes may provide important insights into causal pathways to AD. For example, a recent GWAS demonstrated the effectiveness of this strategy: certain risk genes, such as FRMD6, were first identified to be associated with some intermediate neuroimaging phenotypes like hippocampal atrophy (Shen et al. (2014)), and were later validated as related to AD (Hong et al. (2012); Sherva et al. (2014)). More generally, additional regions of interest (ROIs) or even other imaging features may serve as more effective endophenotypes waiting to be identified.

Existing neuroimaging GWASs almost all rely on *manually* extracted or *pre-defined* imaging features as endophenotypes, such as those based on certain ROIs derived from brain atlases (Shen et al. (2010); Zhang et al. (2014); Zhao et al. (2019, 2021); Zhu et al. (2022)). However, due to limited knowledge about brain structure and function, there is still ongoing debate on how to best define ROIs or even brain atlases; furthermore, these ROIs may or may not be optimally relevant for the given GWAS trait, such as AD in this study (Dickie and et al. (2017)). A motivating question arises here: there are 66 existing atlases for the whole brain structural MRI data (Dickie and et al. (2017)): which one to use? To address these challenges, it would be of high value to develop and apply data-driven methods for novel feature extraction, especially given recent developments and successes in deep learning for image analysis. Due to the high dimensionality of MRI scans, effective feature extraction is both critical and challenging. One possible solution is tensor regression models, along with a scalable algorithm, to reduce the dimension from high-dimensional exposures

(Zhou et al. (2013)). Tensor partition regression models (TPRM) were proposed to enable combining extracted features from partitioned tensors (Miranda et al. (2015)). In recent years, deep learning methods have been increasingly used in feature extraction, especially in neuroimaging. For example, convolutional autoencoder networks have been designed to reduce MRI data dimensionality (Oh et al. (2019); Patel et al. (2022)). Another recent study trained 3D convolutional neural networks to extract features in an AD classification task, achieving an AUC score of 0.75 on the test data and performing GWAS scans to identify genetic variants related to extracted features (Chakraborty et al. (2023)). However, the above methods are not inherently designed for and cannot be interpreted as causal feature extraction, due to the existence of *hidden* confounders affecting both brain images and the AD status (or other outcomes). This limitation may hinder a comprehensive understanding of biological mechanisms underlying AD.

Instrumental variable (IV) regression is a popular method for identifying causal relationships between specific exposures and an outcome while being robust to hidden confounding variables. In the ADNI dataset, exposures are MRI scans or their extracted features, the outcome is the AD status, and SNPs can serve as IVs. The most widely used method in IV regression is two-stage least squares (2SLS) (Klungel et al. (2015)): in Stage 1, it estimates the exposures using IVs, and in Stage 2, the estimated exposures are then used to predict the outcome. In both stages, linear regression models are fitted. When genetic variants/SNPs are utilized as IVs, specific IV regression methods (notably with independent SNPs) are often known as Mendelian randomization (MR), which have recently been applied to neuroimaging data (Mo et al. (2022); Taschler et al. (2022)). One of the most popular MR methods is inverse-variance weighted Mendelian randomization (IVW-MR) (Burgess et al. (2013)), which is designed for a single exposure and combines results from each SNP by inverse-variance weighting. Multi-variable Mendelian randomization

(MVMR) extended this method (IVW-MR) to address multiple exposures (Burgess and Thompson (2015)). To relax the linearity assumption in traditional IV regression, non-parametric models, such as kernel methods and basis functions, have been developed (Hall and Horowitz (2005); Newey and Powell (2003)). For example, a nonparametric method has been proposed to detect a nonlinear causal effect of a single exposure on an outcome using GWAS summary data in transcriptome-wide association studies (TWAS) (Dai et al. (2022a)). Compared to classical non-parametric models, neural networks provide another choice for modeling complicated non-linear relationships. For example, DeLIVR maintained a linear regression model in Stage 1 but used a neural network in Stage 2 (He et al. (2023)), while DFIV (Xu et al. (2020)) and DeepIV (Hartford et al. (2017)) implemented neural networks in both stages to model non-linear causal relationships. DeepGMM (Bennett et al. (2019)), based on the generalized method of moments (GMM), used non-linear functions of exposures and IVs to infer complex causal relationships.

Although existing IV regression methods can extract causal features, there are limitations for some applications, such as those for the ADNI dataset. First, methods like 2SLS and DeLIVR, fitting a linear model in the first stage, are modeling a 3D image by a linear function, which is generally not effective. Furthermore, the ADNI-1 dataset, which is a subset of the ADNI dataset, contains MRI scans of 817 individuals and this number is even smaller when genetic data are combined. Thus, we aim to maximize the use of samples in the ADNI-1 dataset for training and validation, and leverage AD GWAS summary statistics for hypothesis testing. The AD GWAS summary statistics rely on (marginal) linear models between AD status and SNPs. In general, non-linear regression models, which do not ensure a linear relationship between IVs and the outcome, cannot be applied to GWAS summary data. Our models/assumptions are similar to those in a previous study (Dai et al. (2022a)), however, a key difference is that we apply convolutional neural networks (CNNs) to

3-dimensional MRI images as a high-dimensional exposure, while the previous study deals with a scalar (i.e. gene expression) exposure.

To enable application of IV regression to high dimensional exposures and utilization of GWAS summary statistics in testing simultaneously, we propose a novel method called deep feature extraction via instrumental variable regression (DeepFEIVR). For features extracted by a CNN from high-dimensional exposures, DeepFEIVR projects these features onto the space spanned by IVs and utilizes these projected features to predict the outcome, thus ensuring both the relevance/predictivity and causal interpretation of the extracted features with respect to the outcome. In Section 2.2, we describe our proposed novel method DeepFEIVR in detail, as well as its extension for covariate adjustment, DeepFEIVR-CA. Section 2.3 presents simulation results for DeepFEIVR in various settings. In Section 2.4, we apply DeepFEIVR to extract causal features from MRI scans in the ADNI dataset and test their possible associations with AD status using a large-scale AD GWAS summary dataset, followed by a comparison with DeepFEIVR-CA results. Finally, we explore the relationships between the extracted features and specific brain regions or endophenotypes, and conclude with a short discussion.

2.2 Methods

2.2.1 Deep Feature Extraction via Instrumental Variable Regression

2.2.1.1 Causal Model Structure

DeepFEIVR's causal model structure consists of two stages

$$\begin{aligned} \text{Stage 1}: &\quad f_\theta(X) = ZB + U_1 + \epsilon_1, \\ \text{Stage 2}: &\quad Y = f_\theta(X)\beta + U_2 + \epsilon_2, \end{aligned} \tag{2.1}$$

where f_θ is a non-linear multivariate function that extracts q features from X, with $B \in \mathbb{R}^{p\times q}$, $\beta \in \mathbb{R}^q$, and θ estimated in Stage 1 and 2. We assume $q < p$. Confounders U_1 and U_2 are correlated, while ϵ_1 and ϵ_2 are defined as in 2SLS: they are independent and each has a zero mean and constant variance. For simple notation, Z and Y are assumed to be mean-centered at zero. In DeepFEIVR, three IV assumptions are adapted from those in 2SLS by replacing the exposure X with the features $f_\theta(X)$:

1. The conditional distribution of the features $f_\theta(X)$, given Z, is not constant in Z.
2. Z is independent of Y, conditioning on $f_\theta(X)$, U_2 and ϵ_2.
3. Z and $(U_1, U_2, \epsilon_1, \epsilon_2)$ are independent.

The first assumption guarantees that Z is related to $f_\theta(X)$, while the second assumption is directly satisfied if Z, X and Y follow the causal model structure (2.1). Taking the conditional expectation of Y given X yields

$$\mathbb{E}\left(Y|X\right) = f_\theta(X)\beta + \mathbb{E}\left(U_2 + \epsilon_2|X\right),$$

where $\mathbb{E}\left(U_2 + \epsilon_2|X\right) \neq 0$. Ignoring this non-zero term can result in incorrect inference. Thus, IVs are critical here for correct inference about β.

In Stage 1, we can have $\mathbb{E}\left(f_\theta(X)|Z\right) = ZB$, enabling us to project $f_\theta(X)$ onto the space spanned by Z and estimate $\mathbb{E}\left(f_\theta(X)|Z\right)$ by $Z\hat{B}_\theta$. Then, in Stage 2:

$$\mathbb{E}\left(Y|Z\right) = \mathbb{E}\left(f_\theta(X)\beta|Z\right) = \mathbb{E}\left(f_\theta(X)|Z\right)\beta,$$

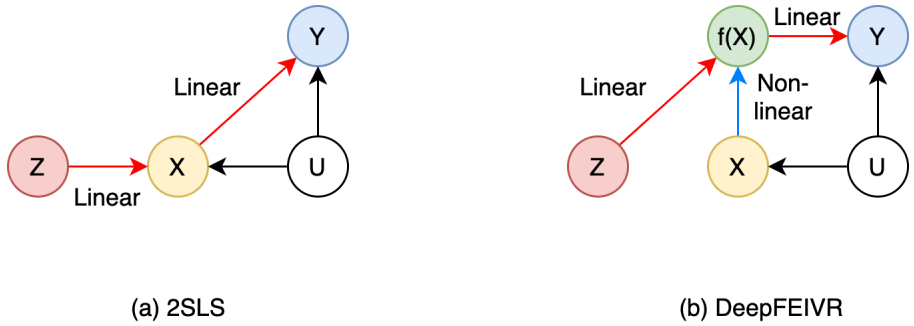

Figure 2.1: Causal model comparison between 2SLS and DeepFEIVR.

in which β can be estimated by replacing $\mathbb{E}(f_\theta(X)|Z)$ with $Z\hat{B}_\theta$ from the previous step. Once θ, B, and β are estimated, $\mathbb{E}(Y|Z)$ can then be estimated by $Z\hat{B}_{\hat{\theta}}\hat{\beta}$, which is a linear transformation of Z. Thus, we can utilize summary statistics in subsequent hypothesis testing within DeepFEIVR, which will be discussed in Section 2.2.1.3.

2SLS can be viewed as a special case of DeepFEIVR by replacing $f_\theta(X)$ with X in the causal model (2.1). Figure 2.1 illustrates the causal models of 2SLS and DeepFEIVR, along with their similarities and differences.

2.2.1.2 Estimation

Aiming to capture the non-linearity of f_θ, we model it as $f_\theta \in \mathcal{F} = \{f_\theta : \theta \in \Theta\}$, in which $\mathcal{F}$ is a class of neural networks with a specified architecture. β and θ are estimated in batches using a stochastic gradient descent (SGD)-type algorithm. For a batch set $\{Z_b, X_b, Y_b\}$ of size n_b, we update the estimates of β and θ by solving

$$\min_{\theta,\beta} \frac{1}{n_b} \|Y_b - Z_b\hat{B}_\theta\beta\|_2^2 + \Omega(\theta, \beta). \tag{2.2}$$

The minimization problem (2.2) handles the second stage. Here, $\Omega(\theta, \beta)$ is an elastic net regularization term for θ and β. In the minimization problem (2.2), $\hat{B}_\theta$ can be obtained by $\hat{B}_\theta = (Z_b^\intercal Z_b + \lambda n_b I)^{-1} Z_b^\intercal f_\theta(X_b)$, which is a closed-form solution for the

following minimization problem in Stage 1:

$$\min_B \frac{1}{n_b}\|f_\theta(X_b) - Z_b B\|_2^2 + \lambda\|B\|_2^2,$$

This is typically a ridge regression model with the ridge penalty parameter λ.

2.2.1.3 Hypothesis Testing

After training the model, we obtain the estimated θ and the estimated weight matrix

$$\hat{B}_{\hat{\theta}} = \left(Z_{tr}^\intercal Z_{tr} + \lambda n_{tr} I\right)^{-1} Z_{tr}^\intercal f_{\hat{\theta}}(X_{tr}), \tag{2.3}$$

which is based on the entire training set. For a global testing of any association between the features $f_\theta(X)$ and the outcome, we test

$$\mathrm{H}_0 : \beta_1 = \beta_2 = ... = \beta_q = 0 \text{ versus } \mathrm{H}_1 : \text{at least one } \beta_j \neq 0.$$

For individual testing of the j-th feature, we test

$\mathrm{H}_0 : \beta_j = 0$ versus $\mathrm{H}_1 : \beta_j \neq 0$.

Next, we demonstrate how hypothesis testing can be conducted based on $\hat{B}_{\hat{\theta}}$ by using an individual-level test dataset $\mathcal{D}_{te}^{indv}$ or summary statistics $\mathcal{D}_{te}^{s}$ derived from $\mathcal{D}_{te}^{indv}$.

Based on an individual-level test set $\mathcal{D}_{te}^{in} = \{Z_{te}, Y_{te}\}$, it is straightforward to perform a global Wald test to assess the association between Y_{te} and $Z_{te}\hat{B}_{\hat{\theta}}$, along with individual Wald tests for each column of $Z_{te}\hat{B}_{\hat{\theta}}$ and Y_{te} to identify (putative) causal relationships between extracted features and the outcome. When the size of the individual-level test set is limited, Wald tests may be inaccurate.

In cases when large individual-level data are unavailable, we can utilize GWAS summary statistics instead, calculated from some large but inaccessible individual-level datasets. Assume that we have summary statistics between Z_{te} and Y_{te} as

$\mathcal{D}^s_{te} = \{\hat{\gamma}_j, \widehat{\text{Var}}(\hat{\gamma}_j)\}_{j=1}^p$ with sample size n_{te}. Following previous studies (Knutson et al. (2020)), we can then estimate the coefficient vector β and its covariance matrix (omitting the test set label for simplicity) as follows:

$$\hat{\beta}_S = \left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z \hat{B}_{\hat{\theta}}\right)^{-1} \hat{B}_{\hat{\theta}}^\intercal Z^\intercal Y \tag{2.4}$$

$$\widehat{\text{Var}}(\hat{\beta}_S) = \hat{\sigma}_S^2 \left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z \hat{B}_{\hat{\theta}}\right)^{-1} \tag{2.5}$$

where

$$\hat{\sigma}_S^2 = \frac{1}{n_{te} - q}\left(Y^\intercal Y - \hat{\beta}_S^\intercal \hat{B}_{\hat{\theta}}^\intercal Z^\intercal Y\right) \tag{2.6}$$

We can approximate $Z^\intercal Z$ by $\frac{n_{te}}{n_R} Z_R^\intercal Z_R$, in which $Z_R \in \mathbb{R}^{n_R \times p}$ is the matrix of Z from a reference panel, such as the ADNI dataset or an independent individual-level genotype dataset. $\{Z^\intercal Y\}_j$ is the j-th component in $Z^\intercal Y$. To estimate $Z^\intercal Y$, we use $\widehat{\{Z^\intercal Y\}}_j = \widehat{\{Z^\intercal Z\}}_{jj}\hat{\gamma}_j$ for each j, in which $\widehat{\{Z^\intercal Z\}}_{jj}$ is the j-th diagonal of the estimated $Z^\intercal Z$. The median of the set $\{(n_{te}-1)\widehat{\{Z^\intercal Z\}}_{jj} \times \widehat{\text{Var}}(\hat{\gamma}_j) + \hat{\gamma}_j \widehat{\{Z^\intercal Y\}}_j, j = 1, 2, ..., p\}$ can be used to approximate $\{Y^\intercal Y\}$. Once $\hat{\beta}_S$ and $\widehat{\text{Var}}(\hat{\beta}_S)$ are obtained, we can perform a global Wald test to test q features globally. For $j = 1, 2, ..., q$, we can test whether the j-th extracted feature is associated with Y, using $\hat{\beta}_{S,j}$ and $\widehat{\text{Var}}(\hat{\beta}_S)_{jj}$. Although Equations (2.4) and (2.5) for $\hat{\beta}_S$ and $\widehat{\text{Var}}(\hat{\beta}_S)$ are derived for a quantitative Y, as discussed in the previous study (Knutson et al. (2020)), they can still be applied to a binary outcome using GWAS summary statistics.

The extracted causal features are not unique. For any invertible matrix $M \in \mathbb{R}^{q \times q}$, the two-stage models can be re-written as: $f_\theta(X)M^{-1} = ZBM^{-1} + U_1 M^{-1} + \epsilon_1 M^{-1}$ and $Y = f_\theta(X)M^{-1}M\beta + U_2 + \epsilon_2$. $f_\theta(X)M^{-1}$ is an equivalent representation of the causal features, with the corresponding association parameter vector $M\beta$. However, this linear transformation does not impact the testing of associations between the extracted features and the outcome because $\beta = 0$ is still equivalent to $M\beta = 0$ for

an invertible matrix M.

2.2.2 DeepFEIVR-Covariate Adjustment (CA)

We extend the proposed method to include covariate adjustment (CA) when certain covariates are present in the training and validation sets. The causal model structure with covariate adjustment is as follows:

$$\begin{aligned} &\text{Stage 1}: \quad f_\theta(X) = ZB + WA + U_1 + \epsilon_1, \\ &\text{Stage 2}: \quad Y = f_\theta(X)\beta + W\gamma + U_2 + \epsilon_2, \end{aligned} \tag{2.7}$$

where W represents w covariates, and $A \in \mathbb{R}^{w\times q}$ and $\gamma \in \mathbb{R}^w$ are the unknown parameters for W. The three IV assumptions in DeepFEIVR can be adapted as: (1) the conditional distribution of $f_\theta(X)$, given Z and W, varies with Z and W; (2) Z is independent of Y conditional on $f_\theta(X)$, W, U_2 and ϵ_2; (3) (Z, W) and $(U_1, U_2, \epsilon_1, \epsilon_2)$ are independent. In Stage 1, we have $\mathbb{E}\left(f_\theta(X)|Z, W\right) = ZB + WA$, and then in Stage 2, $\mathbb{E}(Y|Z, W) = \mathbb{E}\left(f_\theta(X)|Z, W\right)\beta + W\gamma$. Based on the model structure (2.7), we propose DeepFEIVR-CA to include covariates in our model. For parameter estimation in DeepFEIVR-CA using batches, we modify the objective function (2.2) as:

$$\min_{\theta,\beta,\gamma} \frac{1}{n_b} \|Y_b - (Z_b\hat{B}_\theta + W_b\hat{A}_\theta)\beta - W_b\gamma\|_2^2 + \Omega(\theta, \beta), \tag{2.8}$$

in which W_b is the matrix of covariates in a batch set and $\left[\begin{array}{cc} \hat{B}_\theta & \hat{A}_\theta \end{array}\right]^\intercal$ is given by

$$\left[\begin{array}{c} \hat{B}_\theta \\ \hat{A}_\theta \end{array}\right] = \left(\left[\begin{array}{c} Z_b^\intercal \\ W_b^\intercal \end{array}\right]\left[\begin{array}{cc} Z_b & W_b \end{array}\right] + \lambda n_b I\right)^{-1} \left[\begin{array}{c} Z_b^\intercal f_\theta(X_b) \\ W_b^\intercal f_\theta(X_b) \end{array}\right]. \tag{2.9}$$

After training in batches, $\left[\begin{array}{cc} \hat{B}_{\hat{\theta}} & \hat{A}_{\hat{\theta}} \end{array}\right]^\intercal$ is calculated based on the entire training set

by $((Z_{tr}\,W_{tr})^\intercal(Z_{tr}\,W_{tr}) + \lambda n_{tr} I)^{-1}\,(Z_{tr}\,W_{tr})^\intercal f_{\hat{\theta}}(X_{tr})$.

When a large individual-level test dataset including covariates is available, hypothesis testing for β is conducted using an F-test in the model

$$Y_{te} = (Z_{te}\hat{B}_{\hat{\theta}} + W_{te}\hat{A}_{\hat{\theta}})\beta + W_{te}\gamma + \Delta_{te},$$

where Δ_{te} is the error term within the model. However, GWAS summary statistics usually do not provide any information for covariates; assuming that Z and W are nearly uncorrelated (orthogonal), we can proceed with hypothesis testing in DeepFEIVR without covariate adjustment. Under the assumption that Z and W are orthogonal, Equations (2.4) and (2.5) tend to overestimate the covariance matrix of β, producing conservative inference results, as proved in Appendix A.1. In addition, the proof also recommends using $\frac{1}{n_{te}-q-w}\left(Y^\intercal Y - \hat{\beta}_S^\intercal \hat{B}_{\hat{\theta}}^\intercal Z^\intercal Y\right)$ to estimate σ^2 in Equation (2.5) for rigorous results. However, it is important to note that the sample size for summary statistics is typically large, which allows the number of covariates w ($<< n_{te}$) to be neglected.

2.2.3 CNN Model Architecture

In DeepFEIVR, as applied to both the simulated data and the ADNI data, f_θ is a non-linear function estimated by a CNN. In real data analysis, the input to this CNN is 3D MRI images with three channels corresponding to white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF), and the output of this CNN model is the extracted features.

Figure 2.2 illustrates the DeepFEIVR model architecture applied to the simulated dataset, including two CNN layers, a global averaging pooling (GAP) layer, along with two fully connected (FC) layers. Subsequently, we project the q features, which are the output of a leaky ReLU layer, onto the space spanned by IVs to extract q causal features. A final FC layer combines the q extracted features for prediction. CONV

(N@a×a) represents a convolutional layer with N filters and a convolution size of (a, a) and MP (a×a) represents a max pooling layer with a pooling size of (a, a). FC (N) represents a fully connected layer consisting of N neurons. LeakyReLU (a) denotes a leaky ReLU activation with a leakage constant of a.

The left panel of Figure 2.3 presents a CNN module to estimate f_θ both used in the direct CNN model as well as DeepFEIVR, involving four 3D convolutional neural network layers and three fully connected layers. After each CNN layer, we add a 3D max pooling (MP) layer and a batch normalization (BN) layer. A GAP layer is added after the final CNN layer, and two dropout layers are placed between three FC layers. CONV (N@a×a×a) is a convolutional layer with N filters and a convolution size of (a, a, a) and MP (a×a×a) is a max pooling layer with a pooling size of (a, a, a). In the right panel of Figure 2.3, we compare the direct CNN model and DeepFEIVR. In the direct CNN model, a linear regression model is directly applied to the output from the CNN module. In contrast, DeepFEIVR incorporates a projection layer added to the output from the CNN module.

The architecture of DeepFEIVR-CA for the ADNI data is illustrated in Figure 2.4. We project the features extracted by the CNN onto the space spanned by the IVs and covariates. Subsequently, a fully connected layer is applied to a concatenated vector of the projected features and covariates to predict AD.

For a neural network trained on images, Grad-CAM (Selvaraju et al. (2019)) is a technique to identify important image regions from the perspective of a network. In a CNN including a GAP layer, assume $\{Z_{whd} \in \mathbb{R}^{p_1} : w = 1, 2, ..., n_w; h = 1, 2, ..., n_h; d = 1, 2, ..., n_d\}$ represents the output from any layer preceding the GAP layer. Then the activation of voxel (w, h, d) for the j-th feature can be calculated by

$$\left|A^j_{whd}\right| = \left| Z^\intercal_{whd} \left(\frac{1}{n_w n_h n_d} \sum_{whd} \frac{\partial F_j}{\partial Z_{whd}} \right) \right|,$$

in which F_j is the j-th extracted feature. The original version of Grad-CAM, de-

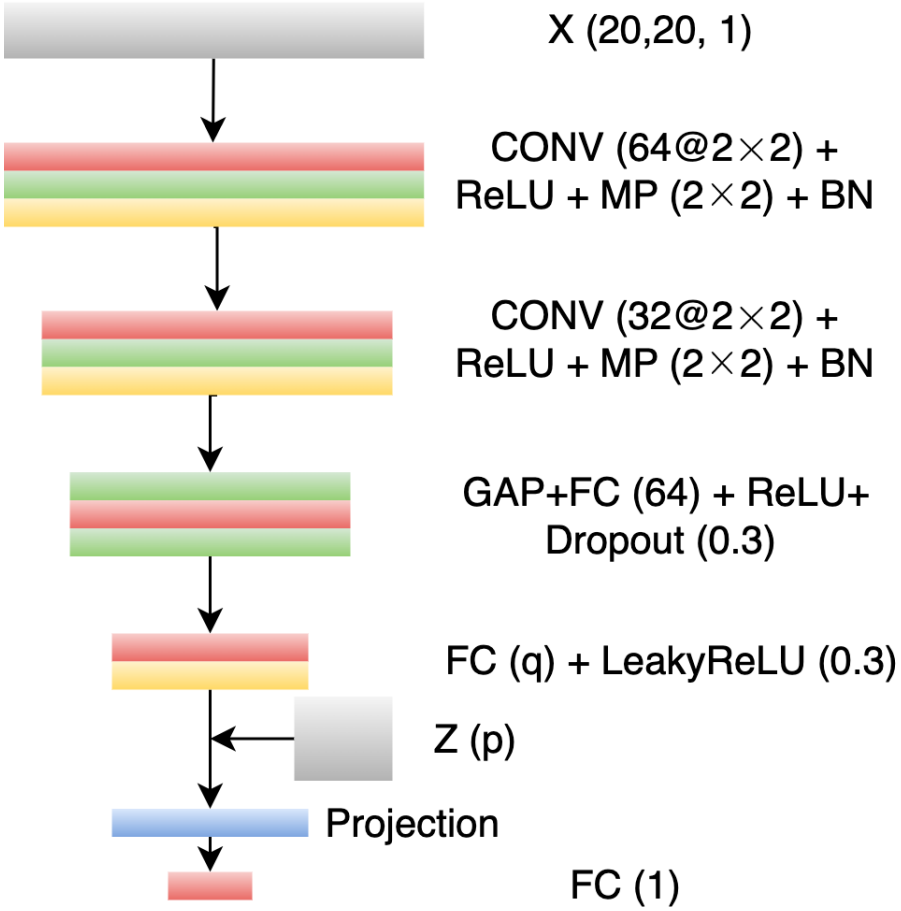

Figure 2.2: The model architecture of DeepFEIVR applied to the simulation dataset.

signed for a classification task, uses $\text{ReLU}(A^j_{whd})$ as the activation function. However, considering that the extracted features in our problem are not binary, we use the absolute value function instead.

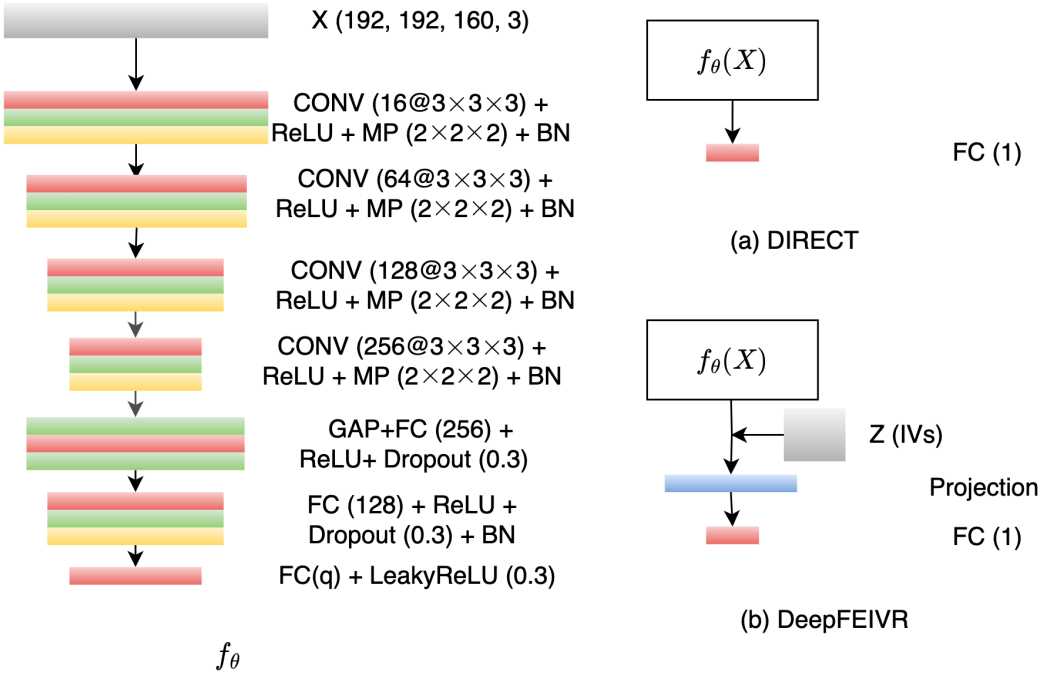

Figure 2.3: Models used for the ADNI dataset. Left: the model architecture of f_θ. Top-Right: the direct CNN model applies a linear regression model following f_θ. Bottom-Right: DeepFEIVR projects the extracted features from f_θ onto space spanned by the IVs, followed by a linear regression model.

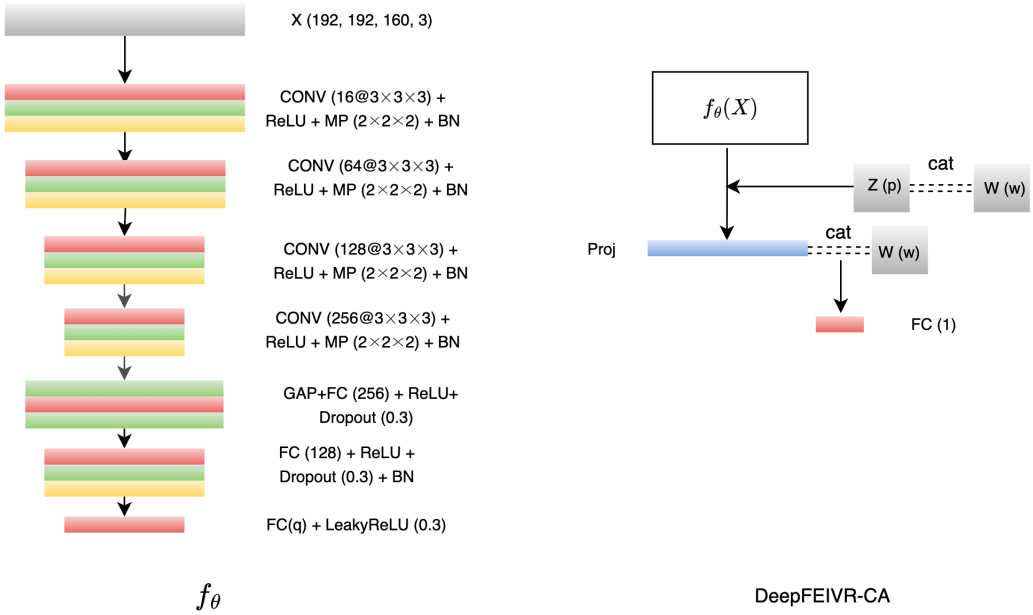

Figure 2.4: The model architecture of DeepFEIVR-CA used in the ADNI dataset. Left: the model architecture of f_θ. Right: DeepFEIVR-CA.

2.3 Simulations

In this section, we assess the performance of DeepFEIVR by conducting a simulation study. In each simulation replicate, we first generate 50 IVs (Z) by $N(0, \Sigma_Z)$ in which Σ_Z is a block diagonal matrix consisting of 10 compound symmetric matrices $\Sigma_Z^{blk} \in \mathbb{R}^{5\times5}$. The off-diagonal elements of Σ_Z^{blk} are 0.1 and the diagonal ones are 1. U and ϵ_2 are independently simulated from $N(0, 1)$, and ϵ_{11} and ϵ_{12} are also independently simulated from $N(0, 0.25)$. Next, X, $F := f_\theta(X)$ (two features) and Y are generated as follows:

$$
\begin{aligned}
F =& \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} = ZB + \begin{pmatrix} U \\ U \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \end{pmatrix}, \\
X =& \text{IMG}(F_1, F_2) + N, \\
Y =& F\beta + U + \epsilon_2,
\end{aligned}
$$

in which the elements in $B \in \mathbb{R}^{50\times2}$ are independently simulated from $N(0, 1)$, and $\beta = d \cdot (-0.1, 0.2)^\intercal$, with d taking values from $\{0.00, 0.02, 0.03, 0.05\}$. The image $X \in \mathbb{R}^{20\times20}$ is generated by an image generating function IMG with the input of two features F_1 and F_2, as well as a noise matrix N of the dimension 20×20, with each element following $N(0, 0.01)$ independently. IMG presents two squares in the left and right parts of the image with values $\sqrt{|F_1|}$ and $\sqrt{|F_2|}$ respectively; the signs of F_1 and F_2 determine whether the squares are placed in the top or bottom parts of the image: if F_1 is positive, the square corresponding to F_1 is placed in the top-left part of the image; otherwise, the square is placed in the bottom-left part. The square corresponding to F_2 is determined similarly but placed in the right part of the image. The sizes of these two squares are independently sampled from three choices $\{4 \times 4, 6 \times 6, 8 \times 8\}$ for some randomness. Figure 2.5 shows an example image (X).

Based on the construction discussed above, F can be interpreted as a function of X, represented by $f_\theta(X)$.

Each simulation setup is repeated 100 times, and in each replicate, we apply DeepFEIVR and 2SLS using both individual-level data and summary statistics. The direct CNN model is also implemented for comparison. In the direct CNN model, the output of the second last layer in the CNN is used as the extracted features and then they are utilized for hypothesis testing. In 2SLS, we regress vec(X), which is the vectorized form of all elements of X, on Z in Stage 1 and conduct hypothesis testing to test the association between Y and $Z\hat{B}_{\text{vec}(X)}$, which is the estimated mean of vec(X). In Stage 2, due to the high dimensionality of vec(X), the degrees of freedom of the Wald test are estimated by the effective rank of $Z\hat{B}_{Vec(X)}$. For DeepFEIVR, we do *not* assume prior knowledge of the true feature number, which is 2 in the simulated data. Therefore, in the standard simulation setting, we set 4 as the (estimated) number of extracted features (q) and use a batch size of 32. In each replicate, the training data size is 800, the validation data size is 200, the individual-level test data size is 4,000, and reference panel size is 20,000. Summary statistics are derived from the same individual-level data in each replicate.

We present the results of the sample proportions of p-values below 0.05 across 100 replicates for DeepFEIVR, the direct CNN and 2SLS in Table 2.1. In addition to the standard simulation setting, the following scenarios are also considered: silencing a feature in the simulations (with $\beta = d \cdot (0.0, 0.2)^\intercal$), utilizing weak IVs (with elements in the first 25 rows of the weight matrix $B \sim N(0, 0.01)$ independently), setting the number of causal features to $q = 2$ (same as the true feature number), using a smaller batch size of 16, and doubling the training sample size (to 1,600).

As presented in Table 2.1, the sample Type I error rates ($d = 0.00$) for DeepFEIVR are close to 0.05, while the sample power ($d = 0.02, 0.03$ and 0.05) grows with larger d. Importantly, the results using summary statistics are close to those using individual-

	Test data	Methods	$d = 0.00$	$d = 0.02$	$d = 0.03$	$d = 0.05$
Standard	Individual	DeepFEIVR	0.05	0.11	0.33	0.74
	Summary	DeepFEIVR	0.05	0.10	0.33	0.74
	Individual	Direct CNN	0.97	1.00	1.00	1.00
	Individual	2SLS	0.08	0.11	0.08	0.28
	Summary	2SLS	0.07	0.10	0.06	0.30
$\beta = d * (0.0, 0.2)^\intercal$	Individual	DeepFEIVR	0.04	0.15	0.32	0.65
	Summary	DeepFEIVR	0.04	0.14	0.33	0.64
	Individual	2SLS	0.01	0.08	0.18	0.24
	Summary	2SLS	0.02	0.07	0.19	0.24
Weak IVs	Individual	DeepFEIVR	0.08	0.08	0.13	0.39
	Summary	DeepFEIVR	0.08	0.07	0.15	0.40
	Individual	2SLS	0.09	0.06	0.09	0.15
	Summary	2SLS	0.10	0.07	0.09	0.22
$q = 2$	Individual	DeepFEIVR	0.05	0.09	0.27	0.70
	Summary	DeepFEIVR	0.05	0.09	0.28	0.70
Batch size 16	Individual	DeepFEIVR	0.06	0.13	0.30	0.80
	Summary	DeepFEIVR	0.06	0.12	0.29	0.80
Training size 1600	Individual	DeepFEIVR	0.05	0.13	0.38	0.77
	Summary	DeepFEIVR	0.07	0.12	0.39	0.77

Table 2.1: Simulation results of DeepFEIVR, the direct CNN and 2SLS in simulations: the sample type I error rates (for $d = 0.0$) and sample power (for $d > 0$) at the nominal significance level of 0.05 based on 100 independent replicates for each setup. .

level test data, confirming the effective use of summary statistics for hypothesis testing within DeepFEIVR. In contrast, the direct CNN fails to identify causal features, as expected, because IVs are not utilized in this model to distinguish true causal features from hidden confounding variables. Furthermore, DeepFEIVR consistently outperforms 2SLS under certain scenarios, whether $\beta = d \cdot (-0.1, 0.2)^\intercal$, $\beta = d \cdot (0.0, 0.2)^\intercal$, or in the presence of weak IVs. For DeepFEIVR, varying the batch size or using a different feature number does not make a big influence on performance, while its power grows slightly with a larger training sample size.

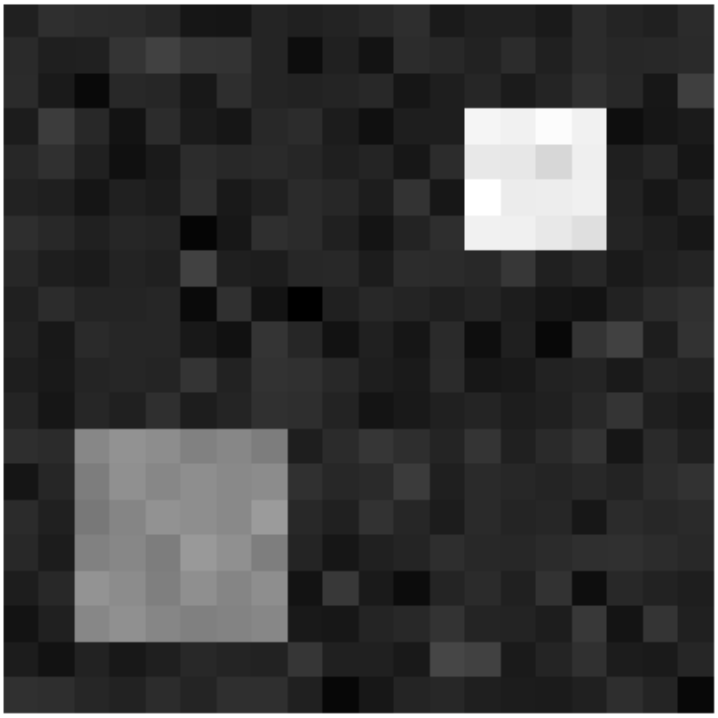

Figure 2.5: A gray scale image example with $F_1 = -1$ and $F_2 = 4$: the square corresponding to F_1 (with dimensions 6×6) is placed in the bottom-left part of the image because F_1 is negative, while the square corresponding to F_2 (with dimensions 4×4) is placed in the top-right part because f_2 is positive.

2.4 Real Data Analyses

2.4.1 Datasets

2.4.1.1 ADNI

In the real data analysis, we use SNPs, MRI scans and AD status in both the training and validation data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (`https://adni.loni.usc.edu`), along with a brain region of interest (ROI) dataset downloaded from ADNI for model interpretation. Established in 2003 and guided by Michael W. Weiner, MD, the ADNI is a collaboration from private and public sources, with the objective of studying the progression of AD by combining MRI scans, genetic data, along with other neuropsychological test results. The ADNI, organized by the Alzheimer's Therapeutic Research Institute at the University of Southern California, is funded by the National Institute of Health, the Department of Defense and numerous other contributors. The complete list of funding sources and organizers for ADNI, along with updated information, can be assessed at `www.adni-info.org`. The original label in the ADNI dataset indicates normal control (CN), mild cognitive impairment (MCI) and AD. In this study, we combine the labels in the MCI and CN categories and define a binary outcome: AD or not AD. 755 individuals in the ADNI dataset have both available MRI scans and SNP data and are included in this study. Among them, 175 individuals are classified as AD. Before applying DeepFEIVR or DeepFEIVR-CA, we select 317 IVs from SNPs by the following steps:

1. We remove SNPs not present in the two summary statistics (IGAP and BIG40) and the reference panel (UK Biobank individual-level genetic data). The details of these datasets are discussed later.

2. SNPs with a missing rate below 20% are retained for further consideration, and missing values are imputed by the mode for each SNP.

3. We fit a linear model to test the association between each SNP and AD status, removing SNPs with a p-value greater than a threshold of 0.001.

4. For highly correlated SNP pairs with the absolute value of correlation greater than 0.8, we only use the SNP with a lower p-value.

After applying GradWarp, B1 non-uniformity, and N3 bias field corrections, ADNI provides T1-weighted 1.5 T 3D structural MRI scans [1]. During MRI scans preprocessing, we first extract brain tissues using the Brain Extraction Tool (BET) (Smith (2002)) and subsequently use the Functional Magnetic Resonance Imaging of the Brain's (FMRIB's) Automated Segmentation Tool (FAST) (Zhang et al. (2001)) to segment tissues into three categories, white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF). Both tools (BET and FAST) are from FSL (FMRIB Software Library). To keep a consistent image dimension from various individuals while preserving important information in modeling the images, we crop all 3D images into the dimension of (192, 192, 130) and normalize voxel values to the range [0, 1]. Each tissue category is treated as a separate channel, and we use the CNN architecture described in Section 2.2.3 to extract features from the 3D images with 3 channels. Figure 2.6 shows a sample image and its three tissues from different planes. In addition to MRI and genetic data, the ADNI dataset also provides an ROI dataset, recording the gray matter volumes of 116 brain regions based on gray matter maps of the ADNI images using longitudinal voxel-based morphometry (VBM) and the regions are determined using anatomical automatic labelling (AAL) (Tzourio-Mazoyer et al. (2002)). This ROI dataset allows us to study how the extracted causal features can relate to specific brain regions.

In the ADNI dataset, 3D MRI scans are treated as the high-dimensional exposure, while a binary variable (AD status) serves as the outcome. Directly extracting

[1] https://adni.loni.usc.edu/data-samples/data-types/mri/

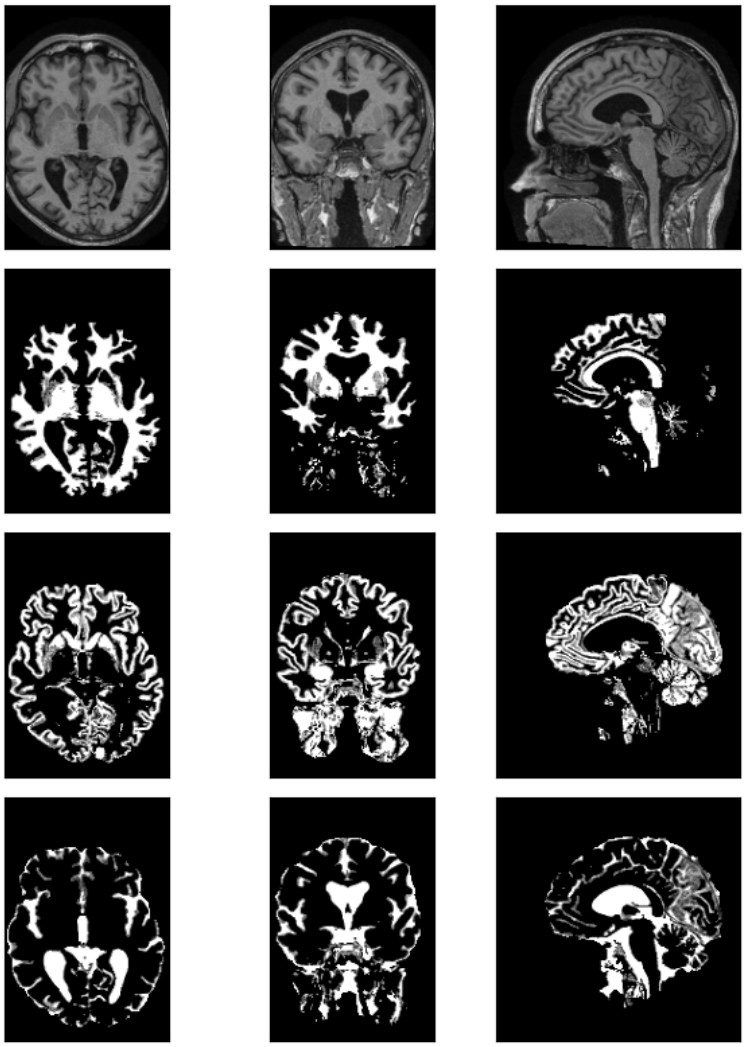

Figure 2.6: A sample image from the ADNI dataset showing the original MRI scan and 3 segmented tissues: WM, GM and CSF tissues (from top to bottom). The axial, coronal and sagittal planes are placed in the left, middle and right. The MRI parameters for this sample image are: a repetition time of 2400 ms, an echo time of 3.5 ms, and a flip angle of 8 degree.

features by training a CNN on MRI scans to predict AD cannot ensure the causality of the extracted features. For example, age can serve as one of the confounders, influencing conditions in brain regions and directly affecting AD risk simultaneously. Changes in brain regions may be strongly correlated with AD, but we cannot assume that these changes are all causal factors. Certain SNPs can be used as IVs since they are likely to satisfy the three IV assumptions: some SNPs are associated with brain structure and function, may affect AD risk only through the brain, and are independent of hidden confounders. In addition, some baseline variables, such as age, gender and handedness, can be used as observed confounders or covariates.

2.4.1.2 IGAP

The limited sample size of the ADNI dataset motivates us to find an additional dataset to test the associations between the extracted features and AD status for higher statistical power. Due to privacy concerns in genetic datasets, large datasets including individual-level genetic and AD status data are not easily accessible. Fortunately, several AD GWAS summary statistics data from large AD GWASs are publicly available. The International Genomics of Alzheimer's Project (IGAP) (Lambert et al. (2013)) is one of them and provides AD summary statistics required for hypothesis testing. IGAP involves two stages of AD GWAS, each providing summary statistics separately. We use the summary statistics from Stage 1, because it contains a larger number of individuals and more SNPs. Specifically, the summary statistics in Stage 1 are calculated based on 17,008 AD cases and 37,154 controls, covering over 7 million SNPs. To avoid identification issues, IGAP only provides the estimated effect sizes of SNPs on AD as well as their standard errors, which are the summary statistics needed in our analysis.

2.4.1.3 UK Biobank Individual-Level Genetic Data

The UK Biobank dataset (Sudlow et al. (2015)) contains individual-level genetic and other data for approximately 490,000 individuals of age 40 and above, collected from 2006 to 2010. We can use this dataset as the reference panel for genotypes in hypothesis testing with GWAS summary statistics, to obtain a robust estimate of $Z^\intercal Z$.

2.4.1.4 BIG40

The Oxford Brain Imaging GWAS Data (BIG40), as a part of the UK Biobank data, records the summary statistics of 3,935 imaging-derived phenotypes (IDPs) based on GWAS performed involving approximately 33,000 individuals in the UK Biobank. For each of 3,935 IDPs and over 17 million SNPs, BIG40 estimates the (marginal) effect size of each SNP and its standard error, serving as summary statistics. The IDPs represent numeric measurements derived from multi-modal MRI images, such as the volume of a specific brain region. We will use the summary statistics to identify significant IDPs associated with the extracted features by DeepFEIVR.

2.4.2 Extracted Features and Associations with Alzheimer's Disease

During the training process of the direct CNN model and the DeepFEIVR models for 672 samples, we use a batch size of 16. The remaining samples are used for validation. In hypothesis testing within DeepFEIVR, we use the IGAP AD GWAS summary data as summary statistics of the test set and the UK Biobank individual-level genetic data as the reference panel. A Wald test is performed on the 20 extracted features, yielding a p-value of 8.321×10^{-11}, which indicates their significant association with AD. We then perform individual tests on each of the 20 features separately. Figure 2.7 shows

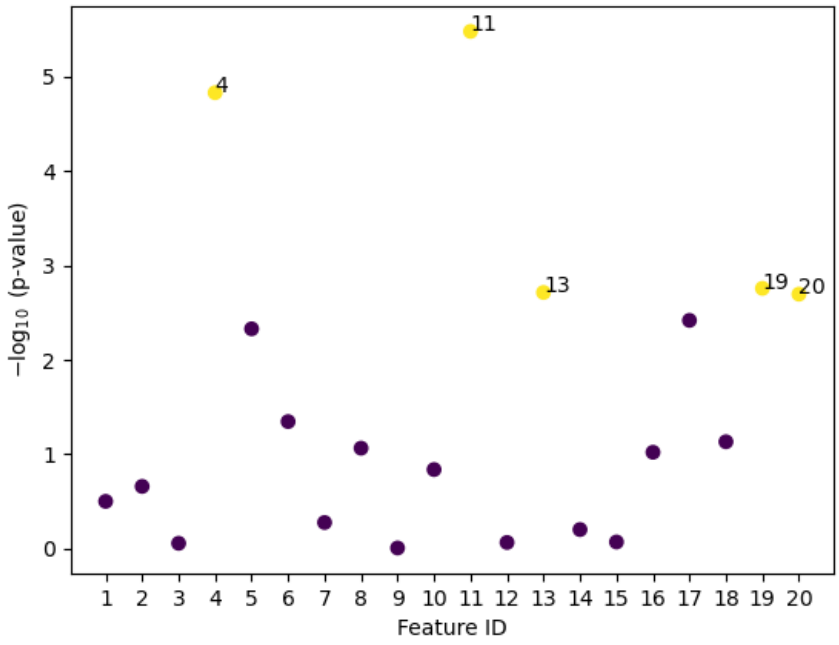

Figure 2.7: Negative $\log_{10}$ (p-values) of individual tests assessing the association between each of the 20 extracted causal features and AD. The extracted causal features with p-values below 0.05/20 are labeled with their IDs.

the $-\log_{10}$ (p-values) of the 20 individual tests, in which the 4th, 11th, 13th, 19th and 20th features are significantly associated with AD at a Bonferroni-corrected p-value threshold of 0.05/20.

As in previous neuroimaging GWASs using the ADNI data (Shen et al. (2010)), we include the baseline age, gender and handedness as covariates for DeepFEIVR-CA. None of the covariates are significantly associated with the IVs, as shown by linear models regressing each covariate on all IVs. The p-value of the global Wald test between the extracted features and AD, based on the IGAP AD GWAS summary statistics, is 0.039. The result is statistically significant but less so significant than that from DeepFEIVR. We will discuss further in the Discussion section of this chapter.

To test how the features extracted by DeepFEIVR and DeepFEIVR-CA are related or different to each other, we perform a canonical correlation analysis (CCA) on the two sets of the extracted features from two methods respectively. CCA stands for a statistical technique to quantify the similarity between two linear subspaces by offering a sequence of maximal correlation coefficients between two sets of orthogonal vectors in the two corresponding subspaces. It provides a non-increasing sequence of CCA correlation coefficients and higher values of the CCA coefficients, especially for the top ones, indicate higher similarity between two features. Specifically, the i-th canonical correlation is the maximum correlation between a linear combination of causal features from DeepFEIVR and a linear combination of causal features from DeepFEIVR-CA, with both combinations are orthogonal to those in the previous $i-1$ canonical correlations. Figure 2.8 presents a canonical correlation plot comparing causal features extracted by DeepFEIVR and DeepFEIVR-CA. The top few components demonstrate a strong relationship: for example, the first four canonical correlations are all exceeding 0.9, although some canonical correlations are moderate or small. For this reason, we will skip further discussions on the extracted causal features from DeepFEIVR-CA.

2.4.3 Interpretation of Extracted Features

2.4.3.1 Extracted Causal Features and Regions of Interest

In this part, we illustrate the relationships between the causal features extracted by DeepFEIVR and the brain regions of interest (ROIs) to facilitate their interpretation. We also compare these causal features with the non-causal features extracted by the direct CNN model (see Figure 2.3 for model differences). To compare 20 features globally, we conduct principal component analysis (PCA) on the causal and non-causal features to extract some top principal components (PCs). The number of PCs

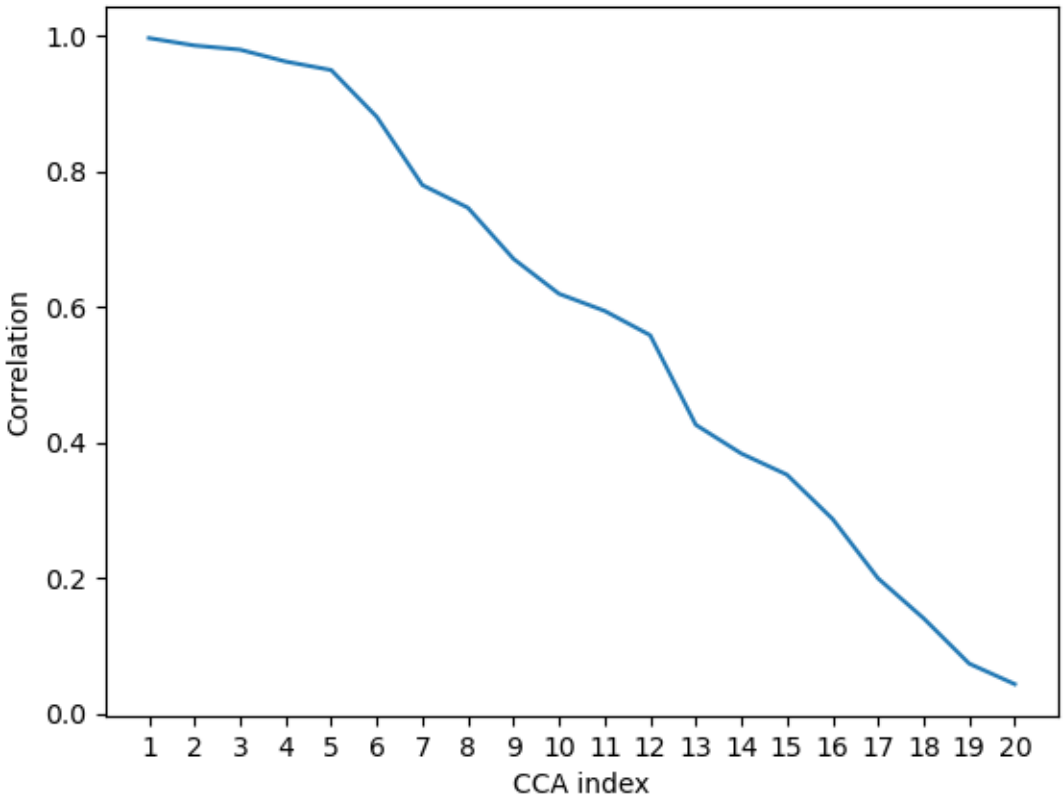

Figure 2.8: The 20 canonical correlation coefficients between extracted causal features from DeepFEIVR and DeepFEIVR-CA.

	Direct CNN	DeepFEIVR
1st	HIPPL	HIPPL
2nd	HIPPR	HESCHLR
3rd	HESCHLR	HIPPR
4th	HESCHLL	TEMPMIDL
5th	ROLANDOPL	TEMPINFL

Table 2.2: The top 5 most significantly associated ROIs for the three noncausal PCs (by the direct CNN) and causal PCs (by DeepFEIVR) after PCA.

is determined based on the criterion that the explained proportion of the total variance by each PC exceeds 10%, because we expect each PC to be representative. For both causal and non-causal features, 3 PCs meet this requirement and cumulatively explain over 70% of the total variance. Then for each ROI and each PC, we fit a simple linear model regressing each PC on each ROI. The p-value of the slope coefficient serves as an indicator of the relatedness between the PC and the ROI. A smaller p-value of the slope coefficient indicates a stronger relationship. Table 2.2 presents the top 5 ROIs most significantly associated with the PCs derived from the features extracted from the direct CNN (non-causal) or DeepFEIVR (causal). The p-value for each ROI is determined by the smallest one among its three p-values across the three PCs.

Table 2.2 presents the top 5 ROIs selected by two methods: left hippocampus (HIPPL), right hippocampus (HIPPR), right Heschl's gyrus (HESCHLR), left Heschl's gyrus (HESCHLL), left Rolandic operculum (ROLANDOPL), left middle temporal gyrus (TEMPMIDL) and left inferior temporal gyrus (TEMPINFL). Among these detected brain regions, the hippocampus has long been known to be associated with AD in the literature (Setti et al. (2017)). The left Heschl's gyrus has been shown to aid in distinguishing AD cases and controls (Hänggi et al. (2011)). In addition, it has been observed that the expression level of γ-aminobutyric acid is affected in the middle temporal gyrus of AD cases (Govindpani et al. (2020)), and a loss of synapses in the inferior temporal gyrus has been reported in individuals with AD or

MCI (Scheff et al. (2011)).

Although there is substantial overlap in the top ROIs associated with causal and non-causal features, the differences still exist between the three PCs from the causal and non-causal features. Canonical correlations between the two sets of the top PCs from the direct CNN and DeepFEIVR are only 0.57, 0.49 and 0.02, suggesting some divergence between the two PC sets.

Figure 2.9 shows the plots of $-\log_{10}$ (p-values) between each of 116 ROIs and each of the 3 causal PCs. For each PC from causal features, significantly associated brain regions are labeled with their names after Bonferroni correction.

Table 2.2 only focuses on the ROIs globally associated with PCs. We then investigate the ROIs related to each feature separately. Figure 2.10 shows p-values for the associations between each of the 116 ROIs and the 13th feature, which is associated with AD based on the IGAP AD GWAS summary statistics. For the 13th feature, HIPPL (left hippocampus), HIPPR (right hippocampus) and HESCHLR (right Heschl's gyrus) are the 3 most significantly associated brain regions and all yield a p-value $< 1 \times 10^{-6}$. However, some significant features detected previously may capture information beyond specific ROIs and are not significantly associated with any ROI.

For comparison, we treat the 166 ROIs as exposures for the outcome AD in multivariable 2SLS, leading to a nearly zero p-value in the global association test. When uni-variable 2SLS is applied to test each ROI and AD, HIPPL, HIPPR and HESCHLR are the most significantly associated ROIs, confirming our previous findings. Note that the ROIs only capture information related to regional gray matter volumes, whereas DeepFEIVR can detect additional brain features, including multiple ROI interactions and characteristics related to other brain tissues, such as white matter.

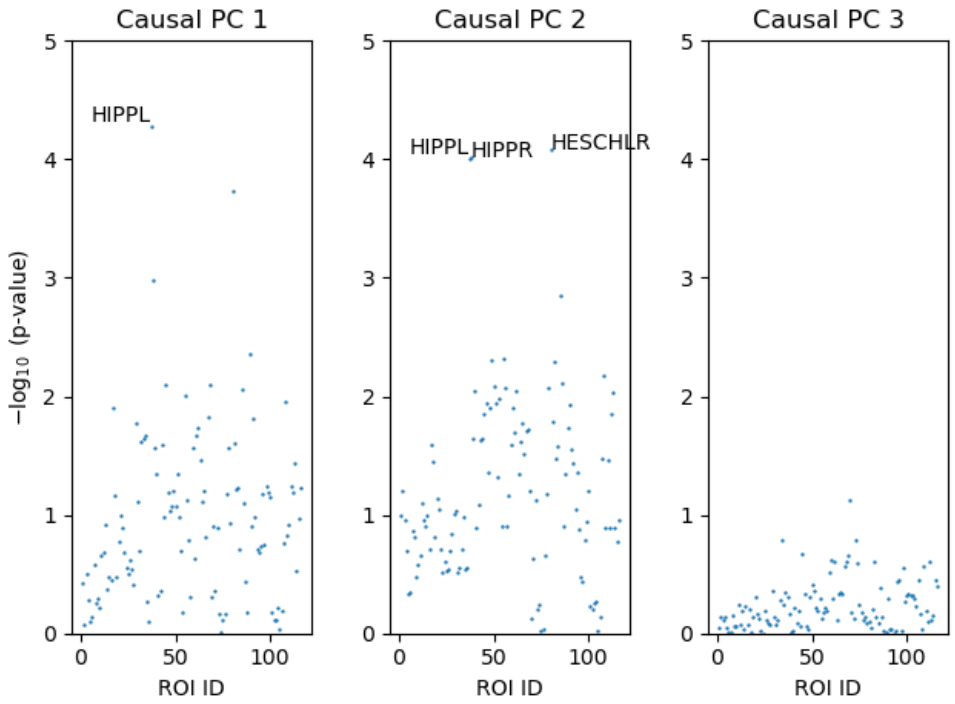

Figure 2.9: Negative $\log_{10}$ (p-values) of association tests between each of 116 ROIs and each of the 3 PCs derived from the extracted causal features.

2.4.3.2 Activation Maps of Extracted Causal Features

In this part, we use Grad-CAM (Selvaraju et al. (2019)) to investigate the relationships between significant features extracted by DeepFEIVR and the brain regions. For each significant causal feature extracted by DeepFEIVR, Grad-CAM can be applied to it and identify the region with the highest absolute activation, which is supposed to contribute the most to that feature. Figure 2.11 presents an example of the activation maps for the 13th feature on the outputs from the second convolutional layer. The absolute activations are scaled to the range [0, 1] in each plot. The detected regions (with high absolute activations) are located near the left and right hippocampus.

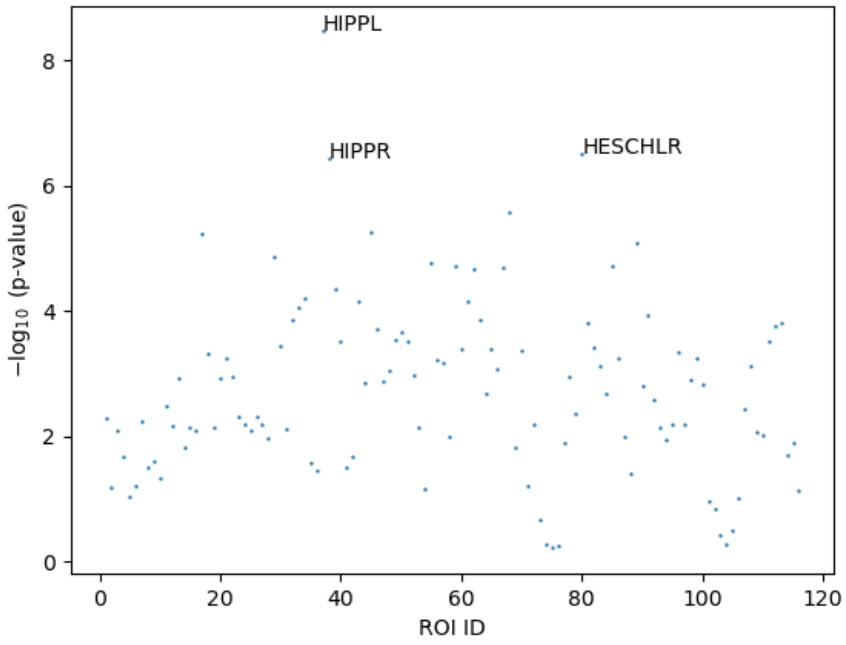

Figure 2.10: Negative $\log_{10}$ (p-values) of association tests between each of 116 ROIs and the 13th extracted causal features.

2.4.3.3 Extracted Causal Features and Imaging-Derived Phenotypes

IDPs are numerical features generated from various types of MRI scans using established methods. BIG40 provides GWAS summary statistics for 3,935 IDPs. For each IDP, we perform a global Wald test on the 20 features extracted by DeepFEIVR. 417 IDPs among these IDPs are marginally significant at a p-value threshold of 0.05, and 114 are significant at a p-value threshold of 0.01. Table 2.3 lists the 10 IDPs with the most significant results. Some IDPs among the 10 selected ones have previously been shown to be related to AD in the literature. For example, individuals with AD exhibit lower fractional anisotropy (FA) and higher mean diffusivity (MD) in regions such as internal capsule and fronto-occipital fasciculus (Mayo et al. (2018)). In addition, cortical thickness is decreased in the lateral occipital cortex for individuals with AD

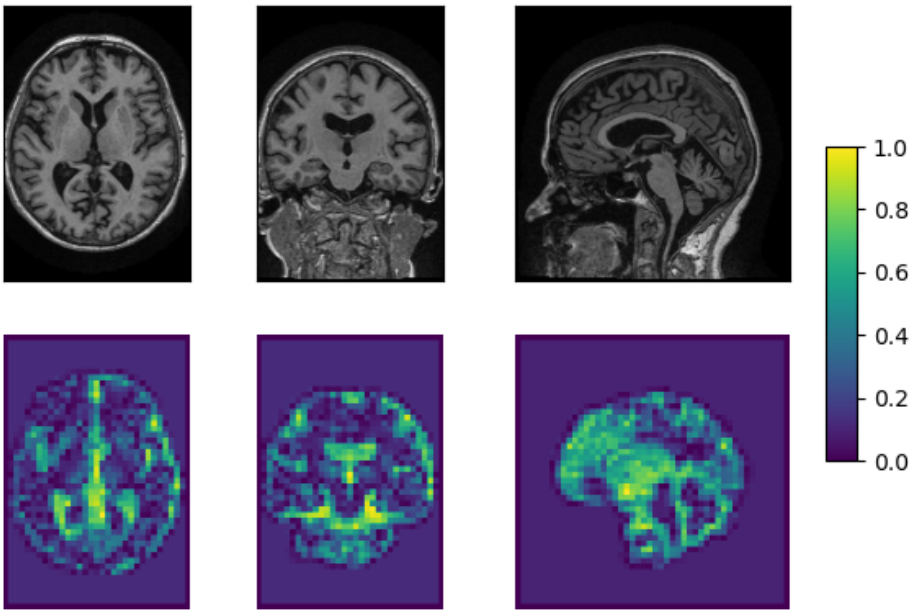

Figure 2.11: Activation maps generated by Grad-CAM for the 13th causal feature.

(Yang et al. (2019)).

Beyond the global Wald testing, we additionally perform individual tests on each of these 10 selected IDPs. Figure 2.12 provides a heat map of $-\log_{10}$(p-values), allowing us to evaluate the relationships between each of the extracted features and the selected IDPs.

IDP (abbr.)	IDP
Vol-P-occ	Volume of Pole-occipital (left) by white surface parcellation (Destrieux)
Vol-Somo	Volume of S-orbital-med-olfact (left) by white surface parcellation (Destrieux)
Area-lo-DK	Area of lateral occipital (left) by white surface parcellation (Desikan-Killiany)
Area-V2	Area of V2 (left) by white surface parcellation (BA_exvivo)
Area-lo-DKT	Area of lateral occipital (left) by white surface parcellation (DKT)
Area-P-occ	Area of Pole-occipital (left) by white surface parcellation (Destrieux)
Area-Soml	Area of S-oc-middle+Lunatus (left) by white surface parcellation (Destrieux)
MD-r-ic	Mean diffusivity in retrolenticular part of internal capsule (right) on fractional anisotropy (FA) skeleton
ML1-sup-fo	Mean L1 in superior fronto-occipital fasciculus (left) on FA skeleton
ML3-r-ic	Mean L3 in retrolenticular part of internal capsule (right) on FA skeleton

Table 2.3: The top 10 most significantly associated IDPs.

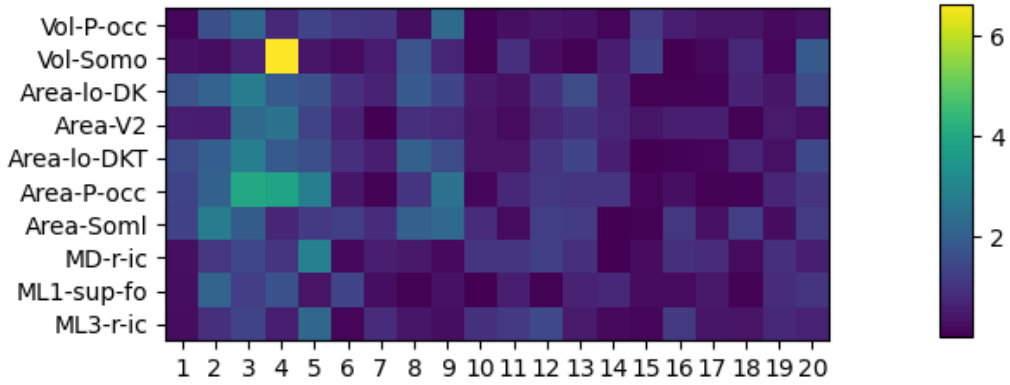

Figure 2.12: The heat map of $-\log_{10}$ (p-values) between each of 20 features and each of 10 IDPs.

2.5 Discussion

In this chapter, we propose a novel deep learning-based instrumental variable regression method called DeepFEIVR to extract causal features from 3D images. A critical benefit of DeepFEIVR is its assumption of a linear relationship between the IVs and the extracted (non-linear image) features, resulting in a linear relationship between the IVs and the outcome. Other distinct advantages of DeepFEIVR include its applicability to high-dimensional exposures, such as 3D neuroimages, for causal feature extraction and to GWAS summary statistics in subsequent feature-outcome association testing. Instead of predicting images directly, DeepFEIVR treats the extracted features as a linear function of IVs during the model training, ensuring that the extracted features are predictive of the outcome and maintain causal interpretation simultaneously. Using the IGAP AD GWAS summary data, we demonstrate that some causal features extracted by DeepFEIVR from the ADNI dataset are significantly associated with AD. By using Grad-CAM, we identify the brain regions most influential to a significantly associated extracted feature in a single input image. In addition, by using brain ROIs and IDPs, we illustrate how the extracted features may be associated with some brain regions or IDPs, thus facilitating the interpretation of the extracted features, although the interpretation remains challenging and largely unsolved for black-box machine learning tools such as CNNs. DeepFEIVR, proposed in this chapter, has broad applicability to other complex biological data, such as ECG data, which will be discussed in Chapter 3.

We also extend DeepFEIVR to DeepFEIVR-CA to allow for covariate adjustment, which can improve the estimation efficiency in the presence of covariates. However, it may not always be necessary because they could be treated as hidden confounders. In our ADNI real data analysis, we have considered and adjusted for the baseline age, gender and handedness, which are often used in previous GWASs in the ADNI data

(Shen et al. (2010)). We found that the DeepFEIVR-CA's extracted features are also associated with AD, though less significant than the extracted features from DeepFEIVR, and that the two feature sets are related. The less significant AD association of the DeepFEIVR-CA-extracted features could be due to (1) hypothesis testing in DeepFEIVR-CA is generally conservative; (2) some covariates are involved in the brain-AD causal pathway. For example, the pathway could be gender/handedness $\Longrightarrow$ brain features $\Longrightarrow$ AD; in such cases, adjusting for these covariates could dilute their effects on some related brain features, thus increasing the risk of missing these brain features in feature extraction. Another issue arises when GWAS summary data lack information about covariates, making hypothesis testing challenging without additional assumptions. In our real data analysis on the ADNI data, we assume that the IVs are (nearly) orthogonal to the covariates, thus enabling us to ignore the covariates in hypothesis testing with GWAS summary statistics, but resulting in conservative inference.

Existing neuroimaging GWAS often rely on manually extracted imaging features or endophenotypes, such as some regions of interest (ROIs) based on a brain atlas. However, due to limited knowledge, there is ongoing debate on how to define ROIs or even brain atlases. With the success of CNNs, especially for automatic feature extraction in image analysis, it is a straightforward step to apply CNNs to extract features from neuroimaging data, and use these features as traits (i.e. endophenotypes) for SNP association analysis (Chakraborty et al. (2023)). CNN-extracted image features could go beyond any pre-defined ROIs, and in this study, we move one step forward: we would like those features to be more likely to be causal to the outcome (i.e. AD) through IV regression. Furthermore, given relatively small sample sizes of existing individual-level neuroimaging genetic datasets, such as the ADNI data, testing extracted features with large-scale GWAS summary statistics, such as the IGAP AD GWAS summary data, can greatly improve statistical power, as demonstrated in our

analyses.

Despite these potential strengths, the proposed approach has some limitations. First, developing a suitable CNN can be time-consuming for a given problem or dataset, especially when determining numerous tuning parameters is required. In our study, a validation dataset aids in improving the predictive performance. Although there is a literature on neural architecture search (NAS) to automate model selection, it remains computationally intensive (Liu et al. (2019)). Second, deep learning models often require large amounts of data. However, the sample size of the ADNI data, around 800, is relatively small. Techniques, such as transfer learning (Dhinagar et al. (2022)), data augmentation (Chakraborty et al. (2023)), and self-supervised learning could be investigated and incorporated to achieve better estimation performance. Third, and perhaps most importantly, the interpretation of CNN-extracted features remains challenging. These topics require future investigations.

2.6 Data and Code Availability Statement

The ADNI dataset is accessible to approved users at `https://adni.loni.usc.edu`. The IGAP summary statistics data can be obtained by approved applicants at `https://www.niagads.org/datasets/ng00036`. The BIG40 dataset is available at `https://open.win.ox.ac.uk/ukbiobank/big40/`. The UK Biobank individual-level genotype data can be downloaded by approved users at `https://www.ukbiobank.ac.uk`. The code to implement our proposed methods is publicly available at `https://github.com/yystat01/DeepFEIVRv1`.

Chapter 3

Extracting Genetically-Imputed Causal Features from ECG Data

3.1 Introduction

Atrial fibrillation (AF), known as the most prevalent form of cardiac arrhythmia, is distinguished by fast and irregular heart rhythms (Ahmed and Zhu (2020); Go et al. (2001)). Epidemiological studies estimate that approximately 2% of the global population is affected by AF, with the prevalence of AF increasing to 10%-17% among those aged over 80 (Zoni-Berisso et al. (2014)). The clinical consequences of AF are substantial. Individuals with AF have a 5 times greater risk for stroke and a 3 times greater risk for death (Chugh et al. (2014); Zoni-Berisso et al. (2014)). The early diagnosis and proactive treatment of AF are important in mitigating these severe outcomes. However, approximately one-third of individuals with AF are asymptomatic, making their AF status frequently undetected (Kornej et al. (2020); Dilaveris and Kennedy (2017)). The asymptomatic nature hinders the detection rate of AF, therefore complicating the effective treatment of AF.

Pulse palpation and blood pressure tests are fast and economical screening tools for AF diagnosis, frequently serving as initial steps before a more definitive diagnosis (Cooke et al. (2006); Dilaveris and Kennedy (2017)). Electrocardiography (ECG) is

the gold standard for the diagnosis of AF (Dilaveris and Kennedy (2017)). In data-based approaches for AF classification, interpretable features have been constructed from ECG recordings and methodologies including support vector machine (SVM) and gradient boosting trees have been applied for the classification of AF (Rizwan et al. (2018); Zheng et al. (2020)). The interpretable features in these studies include summary statistics of ECG signals and durations of the QRS complex. Although predefined features are easy to interpret, their reliance on prior knowledge may neglect underlying details in ECG signals. Beyond interpretable data-based features, model-based features have become more commonly used. Signal processing techniques, such as wavelet transformations, along with time series models, such as autoregressive models, have been employed in extracting features from ECG recordings (Mahmoodabadi et al. (2005); Zhao and Zhang (2005)). The progression of deep learning has established a significant milestone in ECG classification and feature extraction. Within the framework of supervised learning, neural networks, such as deep convolutional neural networks (DCNN), bidirectional long short-term memory (BiLSTM) networks, inception networks, and residual networks, have been implemented to classify AF from ECG recordings (Cheng et al. (2021); Raghunath et al. (2020); Ribeiro et al. (2020)). In addition, unsupervised learning methodologies, including contrastive learning such as patient contrastive learning of representations (PCLR) (Diamant et al. (2022)) and variational Autoencoders (VAEs) (Kuznetsov et al. (2020)), have been applied to extract representative features from unlabeled ECG signals. These extracted features can subsequently be applied in transfer learning or other analytical tasks.

Despite considerable efforts directed toward the extraction of features from ECG recordings and other high-dimensional data, studies focusing on inferring the causality of the extracted features remain limited. A promising solution involves the initial extraction of features from ECG signals, followed by the implementation of causal inference tools, including univariate or multivariate instrumental variable regression.

The application of 2SLS on interpretable ECG features has identified the causal influence of the variations in PR intervals on AF (Gajendragadkar et al. (2021)). Furthermore, the causal associations between genetically imputed spQRSTa in ECG signals and conditions including hypertrophic cardiomyopathy or idiopathic dilated cardiomyopathy were investigated using Mendelian randomization-Egger (MR-Egger) (Young et al. (2023)). However, no evidence of causality was identified in the study. An alternative choice is to implement instrumental variable regression in a high-dimensional framework. DeepFEIVR, as discussed in Chapter 2, is such an approach, leveraging instrumental variable regression to directly extract causal features from high-dimensional data while maintaining the linear association between IVs and outcomes. DeepFEIVR is designed to extract causal features from high-dimensional data by employing genetic variants as IVs. It facilitates the hypothesis testing for the extracted causal features, utilizing solely GWAS summary statistics.

In this chapter, we propose the residual inclusion version of DeepFEIVR, named DeepFEIVR-residual inclusion (DeepFEIVR-RI), aiming at reducing variance in estimating weights compared to DeepFEIVR. A detailed explanation of DeepFEIVR-RI is provided in Section 3.2. The original version of DeepFEIVR is limited to considering only hundreds of IVs, which may not be enough for effective causal inference. To utilize a substantial number of SNPs as IVs, we consider assuming an independent relationship among IVs across linkage disequilibrium (LD) blocks or employing polygenic risk scores (PRS) within LD blocks as IVs to reduce the dimensionality. Subsequently, in Section 3.3.1, we focus on a comparative analysis of the results from the applications of DeepFEIVR and DeepFEIVR-RI to the UK Biobank data (Sudlow et al. (2015)) based on various IV choices. Section 3.3.2 presents contribution maps using the dnn-locate (Dai et al. (2022b)) algorithm to analyze the relationships between ECG components and the outcome.

3.2 Materials and Methods

3.2.1 Causal Model

Based on DeepFEIVR and 2SRI, we propose using the following causal model structure to extract causal features from the high-dimensional exposure:

$$\text{Stage 1: } f_\theta(X) = ZB + U,$$
$$\text{Stage 2: } Y = f_\theta(X)\beta + U\alpha + \epsilon,$$

where U is the hidden confounder and ϵ is an independent noise in Stage 2. A parametric non-linear function f_θ is used to extract q features from the exposure X. θ, β, B, and α are parameters to be learned. Z are IVs satisfying three conditions: (1) Z and $f_\theta(X)$ are correlated; (2) Z can only affect Y through $f_\theta(X)$; (3) Z is independent of (U, ϵ). For the sake of simple notations, it is assumed that $f_\theta(X)$, U, and Z are centered at mean 0.

Since the confounder U affects both X (or $f_\theta(X)$) and Y, ignoring the confounder and fitting $Y = f_\theta(X)\beta + \epsilon'$ in general leads to a biased estimate of the coefficient β, as the new error term $\epsilon' = U\alpha + \epsilon$ and the features $f_\theta(X)$ are dependent. With the inclusion of IVs in DeepFEIVR, β can be instead estimated via $\mathbb{E}(Y|Z) = \mathbb{E}(f_\theta(X)|Z)\beta$, which eliminates terms containing the confounder.

3.2.2 DeepFEIVR-Residual Inclusion (RI)

DeepFEIVR-RI is proposed based on the following working model in Stage 2:

$$Y = f_\theta(X)\beta + \left(f_\theta(X) - Z\hat{B}_\theta\right)\alpha + \epsilon,$$

where $\hat{B}_\theta$ is the estimate of B from Stage 1. The residuals in Stage 1, $f_\theta(X) - Z\hat{B}_\theta$, are employed to estimate the hidden confounder U. When implementing DeepFEIVR-RI

in batches, we replace the minimization problem (2.2) in Chapter 2 with the following task:

$$\min_{\theta,\beta,\alpha} \frac{1}{n_b} \|Y_b - f_\theta(X_b)\beta - \left(f_\theta(X_b) - Z_b \hat{B}_\theta\right)\alpha\|_2^2 + \Omega(\theta; \lambda_1, \lambda_2).$$

DeepFEIVR-RI extends the (linear) two-stage residual inclusion (2SRI). In contrast to the error term $U\alpha+\epsilon$ in Stage 2 of DeepFEIVR, the error term in DeepFEIVR-RI is reduced to ϵ after estimating $U\alpha$ with $\left(f_\theta(X) - Z\hat{B}_\theta\right)\alpha$. Therefore, compared to DeepFEIVR, DeepFEIVR-RI may be able to reduce the variance in estimating β. There is a non-identifiability issue with the parameters. Specifically, scaling $f_\theta(X)$ and β by a constant C and its reciprocal $1/C$ respectively leaves the product $f_\theta(X)\beta$ the same. To mitigate this identification issue, we standardize each feature/component in $f_\theta(X)$ to have a sample mean of 0 and a sample variance of 1.

When obtaining $\hat{\theta}$ in batches, we can estimate B by $\hat{B}_{\hat{\theta}}$ based on all the training samples using Equation (2.3). The inference of extracted features from DeepFEIVR-RI uses the same formulas as presented in Equations (2.4) and (2.5).

3.2.3 DeepFEIVR-RI-CA

In Chapter 2, a covariate adjustment version was developed, named DeepFEIVR-CA, considering adjusting for some covariates or observed confounders, in both stages. Similarly, we extend DeepFEIVR-RI to include covariate adjustment, named DeepFEIVR-RI-CA. The causal model of DeepFEIVR-RI-CA is as follows:

$$\text{Stage 1: } f_\theta(X) = ZB + WA + U,$$
$$\text{Stage 2: } Y = f_\theta(X)\beta + W\gamma + U\alpha + \epsilon,$$

where A and γ are the regression coefficients/weights for covariates W in Stage 1 and 2 respectively. A and B are estimated in Stage 1 using Equation (2.9). Then the estimate of θ and β can be obtained in Stage 2 by

$$\min_{\theta,\beta,\alpha,\gamma} \frac{1}{n_b} \|Y_b - f_\theta(X_b)\beta - \left(f_\theta(X_b) - Z_b\hat{B}_\theta - W_b\hat{A}_\theta\right)\alpha - W_b\gamma\|_2^2 + \Omega(\theta; \lambda_1, \lambda_2).$$

When individual-level test data including covariates is available, as in Section 2.2.2, an F-test can be performed between Y_{te} and $\left(Z_{te}\hat{B}_{\hat{\theta}} + W_{te}\hat{A}_{\hat{\theta}}\right)$, adjusted for W_{te}. When only summary statistics are accessible, and information about covariates is not provided, as in DeepFEIVR, hypothesis testing using summary statistics can only be performed under the assumption that the IVs and the covariates are (nearly) uncorrelated. Under such an assumption, and neglecting the covariate terms, we can use the same hypothesis testing steps as in DeepFEIVR-RI without covariate adjustment.

3.2.4 Instrumental Variable Development

In this chapter, we explore two choices of IVs for genetic association studies: individual SNPs and block-based polygenic risk scores (PRS-blks).

1. Individual SNPs: SNPs in the UK Biobank with a missing data rate exceeding 0.2 and a minor allele frequency (MAF) below 0.01 are excluded. Subsequently, SNP clumping is implemented using a p-value threshold of 0.01, a distance threshold of 250 kilobases (kb), and a linkage disequilibrium (LD) criterion of $r^2 < 0.5$. To simplify computation, we assume that IVs are independent if they are from different IV blocks.

2. PRS-blks: PRS-blks are block-specific PRSs, constructed separately with only SNPs in each LD block utilizing the PRS-cs algorithm (Ge et al. (2019)), lever-

aging summary statistics from an AF GWAS. The data used in the AF GWAS should be independent of data in training and testing for DeepFEIVR and DeepFEIVR-RI.

Before applying DeepFEIVR or DeepFEIVR-RI, we first pre-train the network (without projection and residual inclusion parts) on an ECG AF classification task using the training set. For both types of IVs, individual SNPs and PRS-blks, we perform F-tests on the associations between each IV and the features extracted by the network (without projection) to select strong IVs.

3.2.5 Data

3.2.5.1 UK Biobank and FinnGen Data

We use data from the UK Biobank to analyze the causal effect of ECG signals on AF. The UK Biobank includes approximately 500,000 individuals from 22 centers in the United Kingdom, recruited starting from 2006 (Sudlow et al. (2015)). More information about the UK Biobank can be found on `www.ukbiobank.ac.uk`. After screening the individuals with AF status and good-quality ECG recordings and SNPs, we apply our models (DeepFEIVR and DeepFEIVR-RI) to 44,662 British white individuals in visit instance 2 (beginning 2014). The individuals with good quality of SNPs are defined by the UK Biobank data field 22020, from which related individuals up to the 3rd degree are removed. The number and proportion of AF cases are 1801 and 4.03% respectively. These individuals possess SNPs (as IVs Z), ECGs (as the high-dimensional exposure X), and AF status (as the outcome Y) concurrently. AF status is defined as a combination of self-reported AF from all four visits (starting from 2006, data field 20002 in the UK Biobank) and hospital diagnoses of AF (ranging from 1992 to 2022, data field 41270 in the UK Biobank).

To test the statistical significance of extracted features, we use AF GWAS sum-

mary statistics from FinnGen, a collaborative study involving Finnish biobanks, academic organizations, and global corporate collaborators (Kurki et al. (2023)). In the FinnGen study, the age ranges from 18 to 90 with a median of 63. 98.4% of individuals in the FinnGen study belong to the Finnish ancestry. AF GWAS summary statistics of the FinnGen study are computed based on 45,766 AF cases and 191,924 controls. The proportion of AF cases in the FinnGen study is 19.25%, which is higher than that in the UK Biobank (4.03%). In the real data analysis, we use both the individual-level test set in the UK Biobank and AF summary statistics in the FinnGen study for independent validation.

3.2.5.2 Data Preprocessing

ECG data. In this study, we utilize 12-lead resting ECG recordings in the UK Biobank when individuals are in a calm situation. Each ECG recording was collected within 10 seconds with a sampling frequency of 500 Hz, resulting in a data dimension of $(5000, 12)$. To mitigate issues of severe baseline wandering and heavy noise in the ECG signals, we implement Daubechies 3 Wavelet transform (Daubechies (1992)) and extract the 4th, 5th, and 6th signals from ECG signals after transformation. A comparison of ECGs before and after this preprocessing approach is illustrated in Figure 3.1.

SNP data. Regarding individual SNPs, an entirety of 2,339 SNPs were selected by screening and clumping as mentioned in Section 3.2.4. In training, instead of calculating the projection matrix $Z(Z^\intercal Z + \lambda nI)^{-1}Z^\intercal$ across all individual SNPs, we partition these SNPs into 877 blocks. This block partitioning is pre-defined in Berisa and Pickrell (2016), dividing the whole genome into 1703 blocks and 877 of them contain selected individual SNPs. In addition, we assume SNPs across different blocks are independent. For SNPs within the j-th block, denoted as Z_j, we calculate the block-specific projection matrix. Subsequently, we aggregate the projection matrices

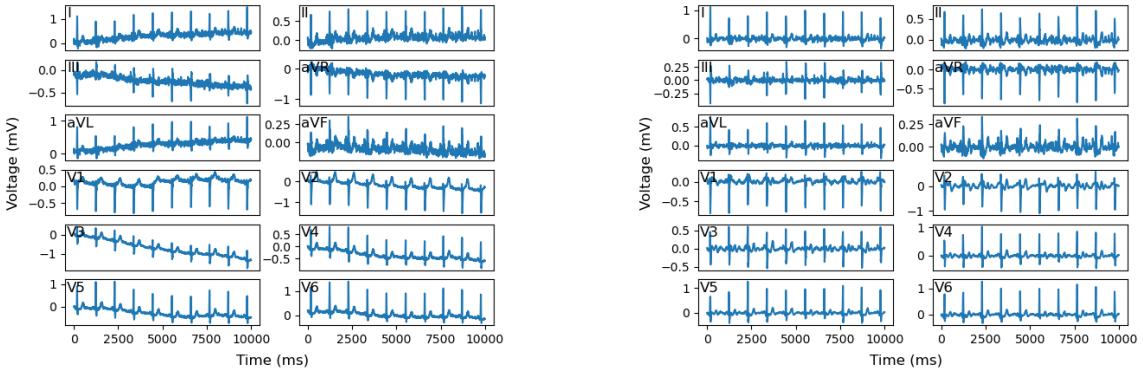

Figure 3.1: An example of the original (left) and preprocessed (right) 12-lead ECG signals in the UK Biobank. In particular, aVR, aVL and aVF are abbreviations for augmented unipolar leads for right arm, left arm and foot.

from all blocks by $\sum_{j=1}^{877} Z_j(Z_j^\intercal Z_j + \lambda n I)^{-1} Z_j^\intercal$. Based on available SNPs in each block, we construct block-based PRS (PRS-blk) using the PRS-cs algorithm with AF summary statistics derived from approximately 310,000 individuals in the UK Biobank using an LD reference panel from European individuals in the 1000 Genomes Project (Auton et al. (2015)). These individuals in the UK Biobank are distinct from those used for training, validating and testing of DeepFEIVR or DeepFEIVR-RI. They possess high-quality genetic data and AF status, but lack ECG data. The proportions of AF cases in two datasets (for AF GWAS summary statistics and for training and inference) are 7.68% and 4.03% respectively. The difference may be due to that some participants in the UK Biobank left the study and their ECG data were not collected in visit instance 2. The characteristics of covariates including age, gender, and handedness are presented in Table B.1 in Appendix B.1. Among the 1703 PRS-blks, 271 are associated with extracted features in the AF classification task and selected as strong IVs.

3.3 Results

3.3.1 Main Results

The dataset is divided into training, validation, and testing subsets with a split ratio of 80%, 10%, and 10%. The architectures of DeepFEIVR-RI and DeepFEIVR-RI-CA are illustrated in Figures 3.2 and 3.3, in which f_θ is derived from a residual network in Ribeiro et al. (2020). Conv1D (F@W, S) represents a one-dimensional convolutional layer with a filter number of F, a window size of W, and a stride size of S (with S=1 omitted as default); MaxPooling1D (P) represents a max pooling layer with a pooling size of P; BN represents batch normalization; FC (N) represents a fully connected layer with N neurons. The network in this AF classification task is initialized by model weights provided in an unsupervised ECG study (Diamant et al. (2022)), and we convert the original dimension $(5000, 12)$ of ECG recordings in the UK Biobank to $(4096, 12)$ through interpolation for the compatibility with the model provided in this unsupervised study. For training DeepFEIVR and DeepFEIVR-RI, we use the Adam algorithm for optimization. Each batch consists of 320 individuals randomly chosen from the training subset. Due to the extreme imbalance of the two classes (with only 4% AF cases in the training set), we assign the weights of 0.96 and 0.04 to AF cases and controls respectively. In training DeepFEIVR and DeepFEIVR-RI using PRS-blks as IVs, we do not standardize IVs because the scale of the IVs measures their potential effects on AF. The number of extracted features is set to be 64, which is a hyper-parameter that can be tuned. We select the number 64 after experimenting with a set of candidates $\{32, 64\}$. According to simulation results in Section 2.3, the number of extracted features has a minor impact on the results. When covariates are considered, we use age, gender and handedness as covariates, in which handedness is coded into three binary predictors: left-handed or not, right-handed or not, and both-handed or not. Individuals with unknown handedness are coded as (0, 0, 0) for

the three binary predictors.

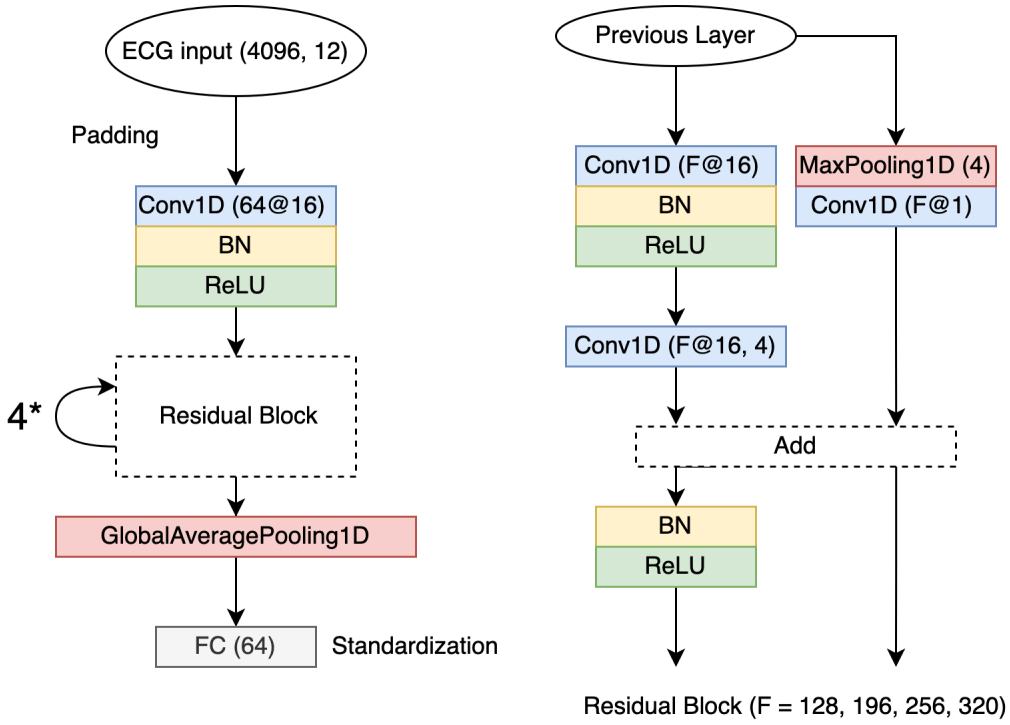

Figure 3.2: The architecture of f_θ in real data analysis.

When covariates are not considered, upon obtaining the estimates of β and θ, we utilize the entire training set $\{Z_{tr}, X_{tr}, Y_{tr}\}$ for the estimation of $\hat{B}_{\hat{\theta}} = (Z_{tr}^\intercal Z_{tr} + \lambda n_{tr} I)^{-1} Z_{tr}^\intercal f_{\hat{\theta}}(X_{tr})$. On the test set $\{Z_{te}, X_{te}, Y_{te}\}$, we compute the AUC scores (along with their corresponding confidence intervals) and the p-values of association tests between the causal feature predictions $Z_{te}\hat{B}_{\hat{\theta}}\hat{\beta}$ and Y_{te}. In addition, we obtain p-values from global Wald tests between causal features $Z_{te}\hat{B}_{\hat{\theta}}$ and Y_{te} using the test set (individual-level data) and the FinnGen AF GWAS summary statistics (Kurki et al. (2023)) with the 1000 Genomes Project as the reference panel. Table 3.1 lists the AUC scores and p-values for DeepFEIVR and DeepFEIVR-RI utilizing individual SNPs and PRS-blks as IVs.

In Table 3.1, for results without covariate adjustment, the performance of Deep-

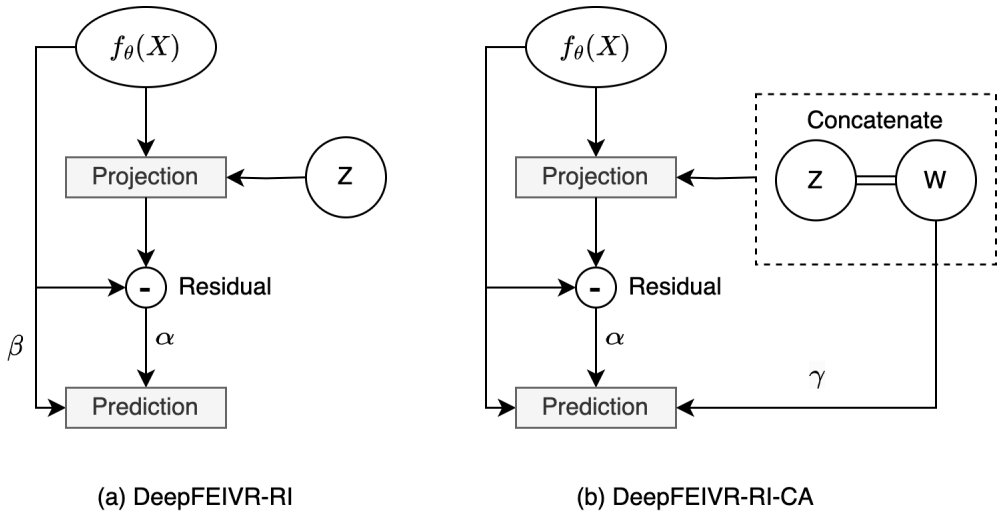

Figure 3.3: The architectures for DeepFEIVR-RI and DeepFEIVR-RI-CA.

FEIVR and DeepFEIVR-RI is close for different choices of IVs. Comparing the performance of individual SNPs in a block structure and PRS-blks, the AUC scores on the individual-level test set using PRS-blks are slightly better than using individual SNPs. Specifically, the confidence intervals of the AUC scores of extracted causal features from DeepFEIVR and DeepFEIVR-RI do not contain 0.5 when using PRS-blks as IVs but the confidence intervals contain 0.5 when using individual SNPs as IVs. The p-values from the Wald tests on their predictions also confirm this finding. The p-values of extracted causal features from the Wald tests on the FinnGen AF GWAS summary statistics are close to 0 for both methods and both IV choices. For both DeepFEIVR and DeepFEIVR-RI, the causal feature predictions and extracted causal features are significantly associated with AF on the individual test set in the UK Biobank using PRS-blks as IVs while not significantly associated with AF using individual SNPs as IVs. The AUC scores of DeepFEIVR and DeepFEIVR-RI with extracted ECG features imputed by PRS-blks are around 0.59. For comparison, the AUC score for the neural network without the projection layer (i.e. genetic imputation) is 0.789 with a CI of (0.751, 0.827). The differences are actually reasonable

because the features extracted by DeepFEIVR and DeepFEIVR-RI after projection are only those associated with genotypes. The aim of this study is to extract causal features from ECGs, then infer the causality between extracted features and AF.

When the covariates are considered, causal feature predictions and causal features are $\left(Z_{te}\hat{B}_{\hat{\theta}} + W_{te}\hat{A}_{\hat{\theta}}\right)\hat{\beta}$ and $Z_{te}\hat{B}_{\hat{\theta}} + W_{te}\hat{A}_{\hat{\theta}}$. The prediction performance of causal feature predictions is much improved for both DeepFEIVR-CA and DeepFEIVR-RI-CA, mainly because the causal features in these models incorporate information from covariates. The procedures for hypothesis testing on causal features for the individual-level test set (with covariates) and the summary statistics (without covariates) are described in Section 2.2.2. Hypothesis testing of causal feature predictions involves an F-test between $\left(Z_{te}\hat{B}_{\hat{\theta}} + W_{te}\hat{A}_{\hat{\theta}}\right)\hat{\beta}$ and Y_{te}, adjusted for covariates W_{te}. For both DeepFEIVR-CA and DeepFEIVR-RI-CA, hypothesis testing of causal feature predictions and causal features from the individual-level UK Biobank test set produces comparable results to DeepFEIVR and DeepFEIVR-RI. The hypothesis testing using the summary statistics from the FinnGen study is based on the assumption that IVs and covariates are independent, and both models yield highly significant inference results. In Table 3.2, we show the p-values of association tests between each covariate and PRS-blks (IVs) in the training set. Only age is marginally associated with PRS-blks (IVs) with a p-value of 0.014 (without adjustment for multiple testing), but age is expected and assumed to be independent of SNPs. The marginal association between age and PRS-blks is likely spurious due to a large training sample size of 35,728. In the following sections, we will only present the results without covariate adjustment.

For DeepFEIVR or DeepFEIVR-RI using individual SNPs as IVs, we intend to include a large number of SNPs to ensure more accurate imputation of ECG features. Although the number of PRS-blks (as IVs) is smaller than the number of individual SNPs (as IVs), PRS-blks may contain more information from more SNPs, which may

Causal feature predictions	IVs	AUC (CI) on UKB Test Data	P-value on UKB Test Data
DeepFEIVR	SNP	0.530 (0.487,0.574)	0.171
DeepFEIVR	PRS-blk	0.589 (0.546,0.632)	6.91×10^{-6}
DeepFEIVR-CA	PRS-blk	0.740 (0.704,0.776)	4.89×10^{-6}
DeepFEIVR-RI	SNP	0.520 (0.476,0.563)	0.349
DeepFEIVR-RI	PRS-blk	0.587 (0.547,0.628)	6.33×10^{-5}
DeepFEIVR-RI-CA	PRS-blk	0.734 (0.698,0.770)	2.36×10^{-5}
Causal features	IVs	P-value on UKB Test Data	P-value on FinnGen Data
DeepFEIVR	SNP	0.180	$< 10^{-8}$
DeepFEIVR	PRS-blk	2.99×10^{-5}	$< 10^{-8}$
DeepFEIVR-CA	PRS-blk	9.77×10^{-5}	$< 10^{-8}$
DeepFEIVR-RI	SNP	0.344	$< 10^{-8}$
DeepFEIVR-RI	PRS-blk	2.06×10^{-5}	$< 10^{-8}$
DeepFEIVR-RI-CA	PRS-blk	7.13×10^{-5}	$< 10^{-8}$

Table 3.1: The AUC scores and their 95% confidence intervals (CIs) for causal feature predictions, the p-values for (linear) associations between causal feature predictions and observed AF statuses based on the individual-level test data in the UK Biobank, and the p-values for (linear) associations of the extracted causal features with (observed) AF based on either the individual-level test data in the UK Biobank or the FinnGen GWAS summary statistics.

	P-value of association tests with PRS-blks (IVs)
Age	0.014
Gender	0.210
Right-handed	0.924
Left-handed	0.912
Both-handed	0.381

Table 3.2: P-values for associations between each covariate and overall 271 PRS-blks (IVs) in the training set.

explain its improved prediction performance. To examine the strength of IVs, we compute the p-values of association tests between extracted causal features before projection and each IV, as shown in Figures 3.4 (individual SNPs) and 3.5 (PRS-blks). In addition, we draw a blue horizontal line indicating a significance level of 0.05. Based on Figures 3.4 and 3.5, 95.90% of individual SNPs and 97.42% of PRS-blks are at least marginally significantly associated with the extracted ECG features before projection.

3.3.2 Model Interpretation

3.3.2.1 Canonical Correlation Analysis (CCA)

In this part, we provide visual comparisons of features extracted by DeepFEIVR-RI using different IVs by canonical correlation analysis (CCA), as discussed in detail in Section 2.4.2. Figure 3.6 showcases the comparison of the extracted causal features using individual SNPs and PRS-blks as IVs, respectively, displaying the top 10 CCA coefficients. With the top 10 canonical correlation coefficients ranging between 0.3 and 0.65, as shown in Figure 3.6, due to different IV choices, the IV-imputed features capture related but different sources of information from the ECG data.

3.3.2.2 Contribution Maps of Block-Based Polygenic Risk Scores or LD Blocks to ECG Causal Features

To help interpret the genetically-imputed (causal) ECG features, we propose using the contribution map of 271 PRS-blks or associated LD blocks to the overall 64 ECG features in Figure 3.7. The contribution score of the j-th PRS-blk to the overall $q = 64$ features follows the path PRS-blk $->$ features, defined as the summation of absolute values of the corresponding weights $\sum_{k=1}^{q} |B_{jk}|$, where B_{jk} is the element of the weight matrix B at the j-th row and the k-th column. In addition to the

contribution score from each PRS-blk, we also provide the contribution score from the j-th corresponding LD block, following the path SNPs (block) $->$ PRS-blk $->$ features, and its contribution score is defined as $\frac{1}{n}\sum_{i=1}^{n}|Z_{ij}|\sum_{k=1}^{q}|B_{jk}|$, where Z_{ij} is the j-th PRS-blk for the i-th individual. $\frac{1}{n}\sum_{i=1}^{n}|Z_{ij}|$ measures the strength of the j-th PRS-blk, and the contribution score from the j-th corresponding LD block measures how much the SNPs in the block contribute to the causal features. The left and right panels in Figure 3.7 present contribution scores for PRS-blks and SNPs in the corresponding blocks. 97.06% of PRS-blks significantly contribute to the causal ECG features with the median of contribution scores divided by 2 as the threshold.

3.3.2.3 Dnn-Loc

Dnn-loc is a data-driven visualization approach providing a statistical interpretation for a neural network (Dai et al. (2022b)). Compared to Grad-CAM used in Chapter 2, dnn-loc is based on a predictive perspective instead of a network perspective. Dnn-loc trains a location network $\delta_\tau(X)$ (under the constraints $\sup_x \|\delta_\tau(X)\|_1 \leq \tau$ and $\sup \|\delta_\tau(X)||_\infty \leq 1$) with the objective

$$\max_{\delta_\tau} \|f_{\hat{\theta}}(X_{tr} - \delta_\tau(X_{tr})X_{tr})\hat{\beta} - Y_{tr}\|_2^2,$$

for features before projection and

$$\max_{\delta_\tau} \|Z_{tr}(Z_{tr}^\intercal Z_{tr} + \lambda n_{tr} I)^{-1} Z_{tr}^\intercal f_{\hat{\theta}}(X_{tr} - \delta_\tau(X_{tr})X_{tr})\hat{\beta} - Y_{tr}\|_2^2,$$

for features after projection where τ is a hyper-parameter controlling the coefficient of determination, either

$$R_\tau^2 = 1 - \frac{\|f_{\hat{\theta}}(X_{tr})\hat{\beta} - Y_{tr}\|_2^2}{\|f_{\hat{\theta}}(X_{tr} - \delta_\tau(X_{tr})X_{tr})\hat{\beta} - Y_{tr}\|_2^2}$$

for extracted (non-causal) features before projection, or

$$R_\tau^2 = 1 - \frac{\|Z_{tr}(Z_{tr}^\intercal Z_{tr} + \lambda n_{tr} I)^{-1} Z_{tr}^\intercal f_{\hat{\theta}}(X_{tr})\hat{\beta} - Y_{tr}\|_2^2}{\|Z_{tr}(Z_{tr}^\intercal Z_{tr} + \lambda n_{tr} I)^{-1} Z_{tr}^\intercal f_{\hat{\theta}}(X_{tr} - \delta_\tau(X_{tr})X_{tr})\hat{\beta} - Y_{tr}\|_2^2}$$

for extracted (causal) features after projection (onto the space of IVs). The idea behind the dnn-locate algorithm is to mask out important locations in ECG recordings that contribute most to the loss function. We select τ from the set $\{2000, 5000, 8000\}$.

In Figure 3.8 for non-causal features (before projection) and Figure 3.9 for causal features (after projection), we present several representative ECG recordings, normalized by their maximum absolute values, along with the detected locations in ECG associated with the extracted features (marked by $\frac{X-\delta_\tau(X)X}{X}$ derived from dnn-loc) by DeepFEIVR-RI using PRS-blks. The detected locations are colored green. For both types of the extracted features in individuals with or without AF, when τ is small, the detected ECG locations are the R waves; as τ increases, the P waves are detected next. R waves are the high peaks in ECG recordings and P waves are the waves on the left side to the R waves.

In addition to ECG recordings, the UK Biobank also provides ECG characteristics, such as R-R intervals. In Table 3.3, we list the p-values of the associations between each ECG characteristic and extracted non-causal features (before projection) or causal features (after projection) by DeepFEIVR-RI using PRS-blks as IVs on the test set. P axis measures atrial depolarization (Acar et al. (2015)) and P onset and offset are the starting and ending points of a P wave respectively. Visualization, such as by dnn-locate, can provide a visual interpretation but only for individual examples, and association tests across all samples in the test set confirm that the extracted non-causal (before projection) and causal features (after projection) from DeepFEIVR-RI are significantly associated with the R and P waves.

3.3.3 Simulations

In simulations, we generally follow the steps in Section 2.3 to assess the performance of DeepFEIVR-RI. The following is repeated for 500 replicates. In each replicate, two features F are generated by $ZB + U$ where $U \sim N(0, 9 \cdot I_2)$ in which $I_2 \in \mathbb{R}^{2\times 2}$

ECG characteristics	Non-causal features	Causal features
R-R interval	$< 10^{-8}$	0.028
P axis	$< 10^{-8}$	0.035
P onset	$< 10^{-8}$	$< 10^{-8}$
P offset	$< 10^{-8}$	$< 10^{-8}$

Table 3.3: P-values of association tests between four ECG characteristics (R-R interval, P axis, P onset, and P offset) and extracted non-causal and causal features by DeepFEIVR-RI.

is an identity matrix. IVs Z and images X are generated under the same settings as in Section 2.3. Additionally, the outcome Y is then generated by $F\beta + U\alpha + \epsilon$, in which $\beta = (-0.1d, 0.2d)^\intercal$ with $d = 0.00$, 0.03, 0.05, and 0.10, $\alpha = (-1, 1)^\intercal$, and ϵ independently follows $N(0, 1)$. Compared to the original simulation settings in Section 2.3, we increase the influence from the hidden confounders.

In each replicate, we implement DeepFEIVR and DeepFEIVR-RI using individual-level data and summary statistics and compare their performance. The network model architecture is the same as the one used in Section 2.3. Training, validation, and testing sizes are 800, 200, and 4,000 and the sample size used to estimate $Z^\intercal Z$ in the test set (reference panel) is 20,000, which are the same as in Section 2.3. In Table 3.4, the sample type I error rates when $d = 0.00$ and sample power when $d = 0.03$, 0.05, and 0.10 are presented. The results confirm that summary statistics are sufficient in hypothesis testing for DeepFEIVR-RI and DeepFEIVR-RI can outperform DeepFEIVR when d is small, especially when $d = 0.03$. This finding is expected, as when d is small, the variation of Y is largely explained by the confounder, under which DeepFEIVR-RI estimates the term related to the confounder ($U\alpha$) to reduce the variance of the error term.

Model	Data	$d = 0$	$d = 0.03$	$d = 0.05$	$d = 0.1$
DeepFEIVR	Individual-level	0.076	0.080	0.160	0.432
DeepFEIVR	Summary statistics	0.078	0.074	0.172	0.428
DeepFEIVR-RI	Individual-level	0.068	0.112	0.170	0.430
DeepFEIVR-RI	Summary statistics	0.064	0.114	0.172	0.436

Table 3.4: Simulation results: sample type I error rates ($d = 0.00$) and power ($d = 0.03$, 0.05, or 0.10) for DeepFEIVR and DeepFEIVR-RI using the individual-level test set or summary statistics.

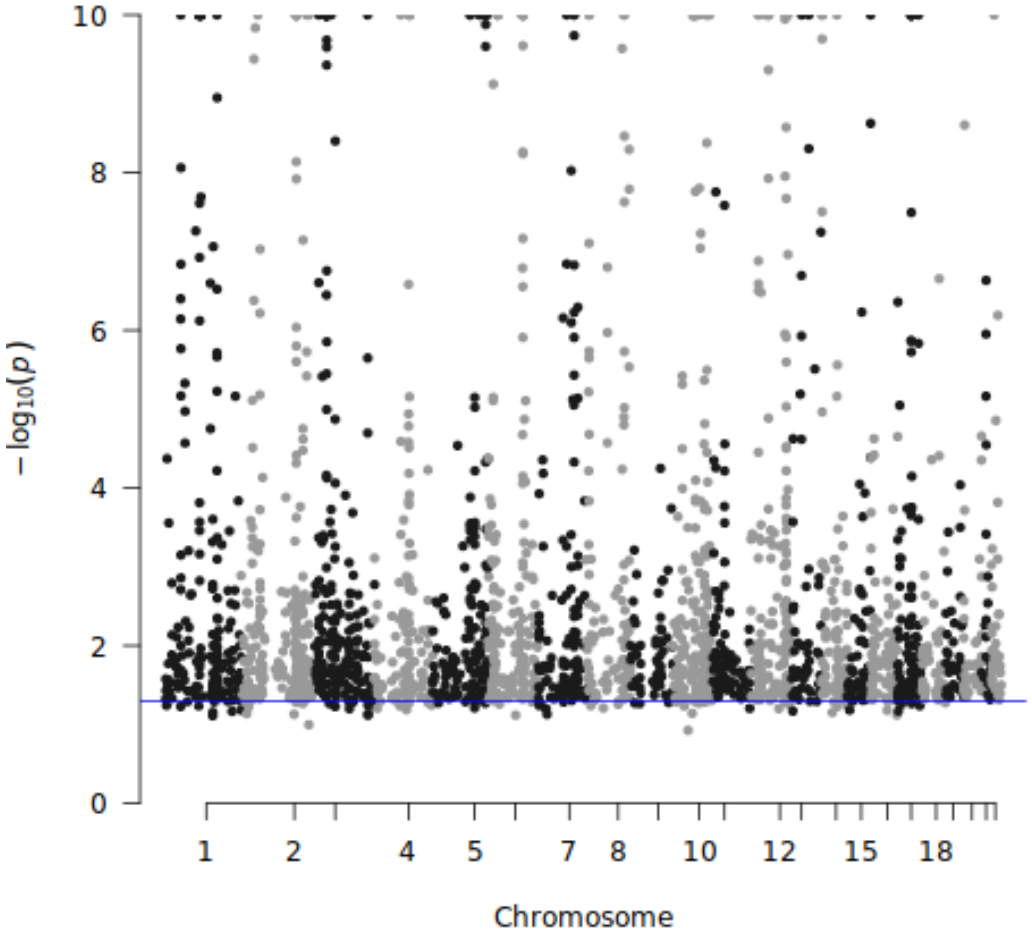

Figure 3.4: Negative $\log_{10}$(p-values) across 64 features for each individual SNP.

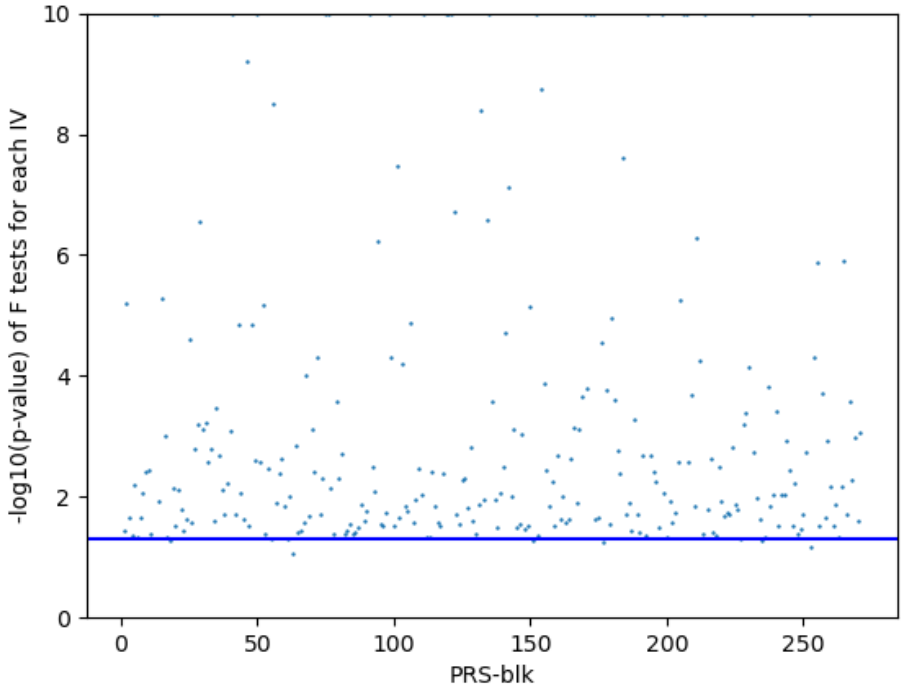

Figure 3.5: Negative $\log_{10}$(p-values) across 64 features for each PRS-blk. The values are truncated at 10.

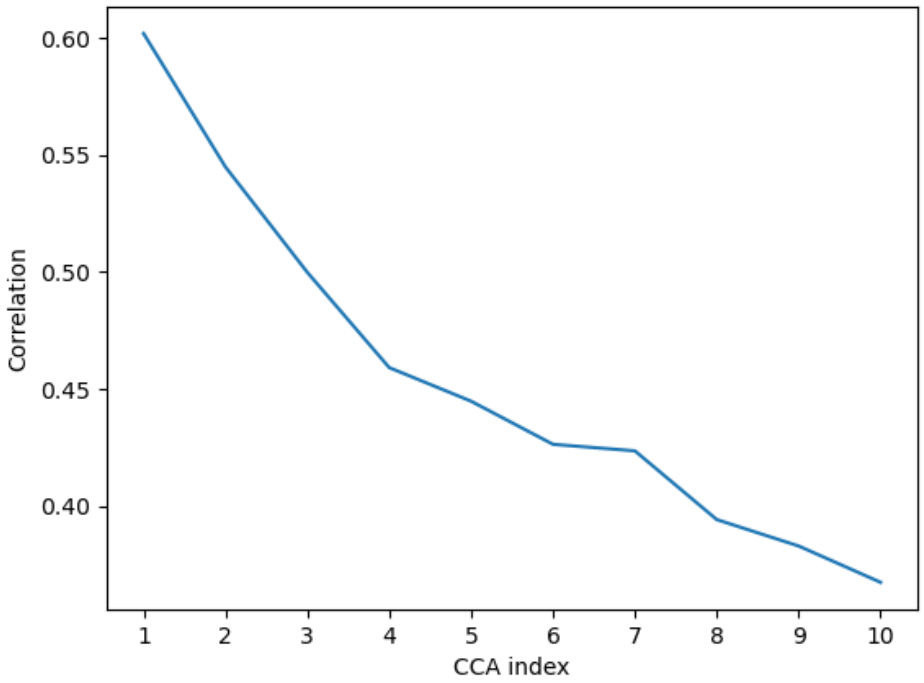

Figure 3.6: Top 10 CCA coefficients between the extracted features using individual SNPs or PRS-blks as IVs by DeepFEIVR-RI.

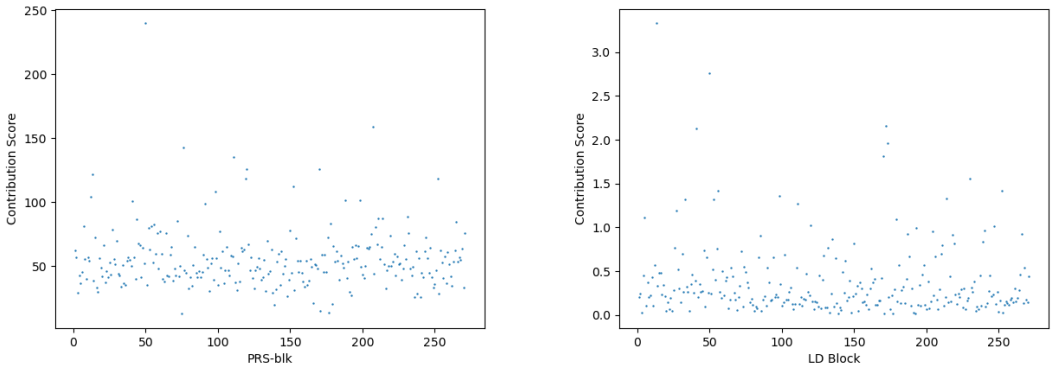

Figure 3.7: The contribution maps of global 64 features from each PRS-blk (left panel), and from each of 271 associated LD blocks (right panel).

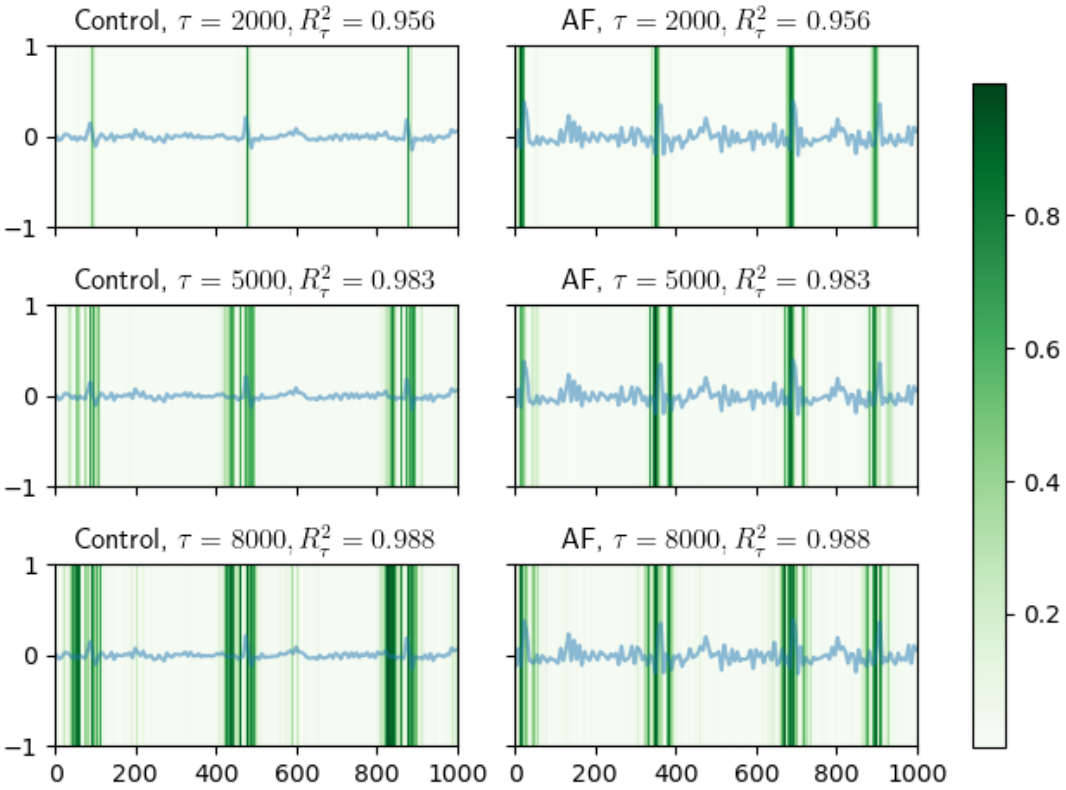

Figure 3.8: Localized important features (before projection) extracted by DeepFEIVR-RI (in green) in example ECG signals (blue, the first 1000 points in lead I) in the test set (left three panels: ECG signals of individuals without AF; right three panels: individuals with AF).

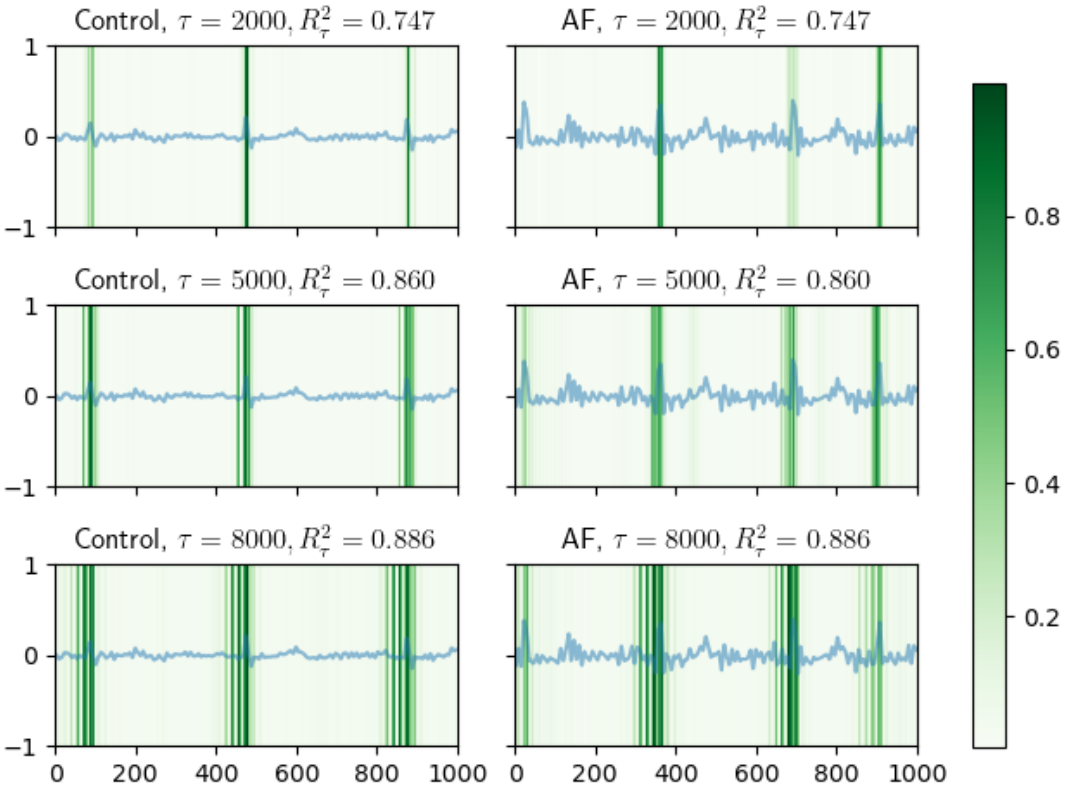

Figure 3.9: Localized important causal features extracted by DeepFEIVR-RI (in green) in example ECG signals (blue, the first 1000 points in lead I) in the test set (left panels: ECG signals of individuals without AF; right panels: individuals with AF).

3.4 Discussion

In this chapter, we have applied DeepFEIVR and its extension, DeepFEIVR-RI, to the UK Biobank dataset to extract causal features of ECG signals associated with AF. Compared to the study in Chapter 2, our contributions include (1) application to a large dataset; (2) development of DeepFEIVR-RI, a residual inclusion extension of DeepFEIVR to reduce estimation variance; (3) construction of IVs from a large number of SNPs; and (4) adaptation of dnn-loc for visual interpretation of causal features extracted by DeepFEIVR-RI. For IV selection, we have explored two distinct approaches: individual SNPs, and block-based PRSs (PRS-blks). We have observed that the AUC scores on the individual-level test data, using PRS-blks, are approximately 0.59 for DeepFEIVR and DeepFEIVR-RI, marginally surpassing the performance using individual SNPs. Note that the extracted causal features are genetically imputed; it is expected that the prediction performance of these extracted causal features will be less competitive, especially compared with direct classification models. The improvement of using PRS-blks over individual SNPs as IVs comes from its ability to incorporate information from numerous SNPs of weak effects while avoiding using a large number of IVs individually. It has been confirmed previously that the utilization of PRSs as IVs can reduce the bias in estimating the causal parameters by avoiding weak IV bias, compared to using individual genetic variants directly as IVs (Burgess and Thompson (2013)). One advantage of DeepFEIVR and DeepFEIVR-RI is their ability to infer causal relationships in test data using solely GWAS summary statistics. The hypothesis testing on the FinnGen AF GWAS summary statistics shows that the genetically-imputed ECG features are indeed associated with AF in FinnGen, even when the proportion of AF cases in the FinnGen study is far different from the proportion in the UK Biobank. In addition, the application of CCA demonstrates both the relatedness and differences between the extracted features based on

the two different IV choices. Application of the dnn-loc algorithm reveals that the R waves and P waves in ECG signals are associated with AF through the extracted causal features by DeepFEIVR-RI, suggesting that the R waves and the P waves are putative causal features for AF. It has been known that AF is dominantly related to irregular R-R intervals (indicated by R waves) and flat P waves, confirming our findings (Ahmed and Zhu (2020); Rasmussen et al. (2020)).

Nonetheless, there are some limitations in this study. First, although we standardize $f_\theta(X)$ to alleviate the parameter identification issue, the estimate of β is not sign-unique. But we note that the sign issue with $f_\theta(X)$ and β does not seem to affect the prediction and hypothesis testing (with a chi-squared null distribution that does not depend on the sign of the test statistic). Second, the improvement of DeepFEIVR-RI over DeepFEIVR in performance is insignificant in terms of their AUC scores, though DeepFEIVR-RI is expected to reduce variance in estimating β and the improvement could be observed under certain scenarios in simulation studies. Finally, we select individual SNPs or PRS-blks associated with AF as IVs. This is based on the simple idea that an IV associated with any causal ECG features for AF is expected to be associated with AF. On the other hand, however, some of these IVs may directly affect AF and thus are likely to be invalid IVs. Due to the nature of the model structure, invalid IVs are difficult to identify. Without assuming valid IVs, as in transcriptome-wide association studies (TWAS) (Gamazon et al. (2015); Gusev et al. (2016)), our method can be interpreted as extracting genetic components of ECG data that are associated with AF, which are expected to be biologically more relevant and more robust to many environmental confounders and experimental artifacts. These issues warrant future studies.

3.5 Data Availability Statement

The UK Biobank data can be accessed for approved users at `https://www.ukbiobank.ac.uk`. The FinnGen AF summary statistics data can be downloaded from `https://www.finngen.fi/`. The 1000 Genomes Project individual-level SNP data is available at `http://www.internationalgenome.org`.

References

Acar, R. D., Bulut, M., Acar, Ş., Izci, S., Fidan, S., Yesin, M., and Efe, S. C. (2015). Evaluation of the p wave axis in patients with systemic lupus erythematosus. *J Cardiovasc Thorac Res*, 7(4):154–157.

Ahmed, N. and Zhu, Y. (2020). Early detection of atrial fibrillation based on ecg signals. *Bioengineering*, 7(1).

Apostolova, L. G. (2016). Alzheimer disease. *Continuum: Lifelong Learning in Neurology*, 22(2 Dementia):419.

Association, A. (2022). 2022 alzheimer's disease facts and figures. *Alzheimers Dement*, 18(4):700–789.

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., and et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Beltran, J. F., Wahba, B. M., Hose, N., Shasha, D., Kline, R. P., and the Alzheimer's Disease Neuroimaging Initiative, F. (2020). Inexpensive, non-invasive biomarkers predict alzheimer transition using machine learning analysis of the alzheimer's disease neuroimaging (adni) database. *PLOS ONE*, 15(7):1–26.

Bennett, A., Kallus, N., and Schnabel, T. (2019). Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32.

Berisa, T. and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283–285.

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665.

Burgess, S. and Thompson, S. G. (2013). Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology*, 42(4):1134–1144.

Burgess, S. and Thompson, S. G. (2015). Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects. *American Journal of Epidemiology*, 181(4):251–260.

Chakraborty, D., Zhuang, Z., Xue, H., Fiecas, M. B., Shen, X., and Pan, W. (2023). Deep learning-based feature extraction with mri data in neuroimaging genetics for alzheimer's disease. *Genes*, 14(3).

Cheng, J., Zou, Q., and Zhao, Y. (2021). Ecg signal classification based on deep cnn and bilstm. *BMC Medical Informatics and Decision Making*, 21(1):365.

Chugh, S. S., Havmoeller, R., Narayanan, K., Singh, D., Rienstra, M., Benjamin, E. J., Gillum, R. F., Kim, Y.-H., McAnulty, J. H., Zheng, Z.-J., Forouzanfar, M. H., Naghavi, M., Mensah, G. A., Ezzati, M., and Murray, C. J. (2014). Worldwide epidemiology of atrial fibrillation. *Circulation*, 129(8):837–847.

Cooke, G., Doust, J., and Sanders, S. (2006). Is pulse palpation helpful in detecting atrial fibrillation? a systematic review: particular high-risk patients may benefit from repeated testing. *Journal of family practice*, 55(2):130–135.

Dai, B., Li, C., Xue, H., Pan, W., and Shen, X. (2022a). Inference of nonlinear causal effects with gwas summary data, arxiv, https://arxiv.org/abs/2209.08889.

Dai, B., Shen, X., Chen, L. Y., Li, C., and Pan, W. (2022b). Data-adaptive discriminative feature localization with statistically guaranteed interpretation. *Annals of Applied Statistics*.

Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM.

Dhinagar, N. J., Thomopoulos, S. I., Rajagopalan, P., Stripelis, D., Ambite, J. L., Ver Steeg, G., and Thompson, P. M. (2022). Evaluation of transfer learning methods for detecting alzheimer's disease with brain mri. *bioRxiv*.

Diamant, N., Reinertsen, E., Song, S., Aguirre, A. D., Stultz, C. M., and Batra, P. (2022). Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLoS Computational Biology*, 18(2):e1009862.

Dickie, D. and et al. (2017). Whole brain magnetic resonance image atlases: A systematic review of existing atlases and caveats for use in population imaging. *Front Neuroinform*, 11:1.

Dilaveris, P. E. and Kennedy, H. L. (2017). Silent atrial fibrillation: epidemiology, diagnosis, and clinical impact. *Clin Cardiol*, 40(6):413–418.

Gajendragadkar, P. R., Von Ende, A., Ibrahim, M., Valdes-Marquez, E., Camm, C. F., Murgia, F., Stiby, A., Casadei, B., and Hopewell, J. C. (2021). Assessment

of the causal relevance of ecg parameters for risk of atrial fibrillation: A mendelian randomisation study. *PLoS Medicine*, 18(5):e1003572.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., Im, H. K., and Consortium, G. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature Communications*, 10(1):1776.

Go, A. S., Hylek, E. M., Phillips, K. A., Chang, Y., Henault, L. E., Selby, J. V., and Singer, D. E. (2001). Prevalence of Diagnosed Atrial Fibrillation in AdultsNational Implications for Rhythm Management and Stroke Prevention: the AnTicoagulation and Risk Factors In Atrial Fibrillation (ATRIA) Study. *JAMA*, 285(18):2370–2375.

Govindpani, K., Turner, C., Waldvogel, H. J., Faull, R. L. M., and Kwakowsky, A. (2020). Impaired expression of gaba signaling components in the alzheimer's disease middle temporal gyrus. *International Journal of Molecular Sciences*, 21(22).

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., Raitakari, O. T., Kuusisto, J., Laakso, M., Price, A. L., Pajukanta, P., and Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252.

Hall, P. and Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904 – 2929.

Hänggi, J., Streffer, J., Jäncke, L., and Hock, C. (2011). Volumes of lateral temporal and parietal structures distinguish between healthy aging, mild cognitive impairment, and alzheimer's disease. *J Alzheimers Dis*, 26(4):719–734.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1414–1423. PMLR.

He, R., Liu, M., Lin, Z., Zhuang, Z., Shen, X., and Pan, W. (2023). Delivr: a deep learning approach to iv regression for testing nonlinear causal effects in transcriptome-wide association studies. *Biostatistics*.

Hong, M., Reynolds, C., Feldman, A., Kallin, M., Lambert, J., Amouyel, P., Ingelsson, E., Pedersen, N., and Prince, J. (2012). Genome-wide and gene-based association implicates frmd6 in alzheimer disease. *Hum Mutat*, 33:521–529.

Klungel, O., Uddin, M. J., de Boer, A., Belitser, S., Groenwold, R., Roes, K., et al. (2015). Instrumental variable analysis in epidemiologic studies: an overview of the estimation methods. *Pharm Anal Acta*, 6(353):2.

Knutson, K. A., Deng, Y., and Pan, W. (2020). Implicating causal brain imaging endophenotypes in alzheimer's disease using multivariable iwas and gwas summary data. *Neuroimage*, 223:117347.

Kornej, J., Börschel, C. S., Benjamin, E. J., and Schnabel, R. B. (2020). Epidemiology of atrial fibrillation in the 21st century. *Circulation Research*, 127(1):4–20.

Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., Reeve, M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Della Briotta Parolo, P., Lehisto, A. A., Kanai, M.,

Mars, N., Rämö, J., and et al. (2023). Finngen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944):508–518.

Kuznetsov, V. V., Moskalenko, V. A., and Zolotykh, N. Y. (2020). Electrocardiogram generation and feature extraction using a variational autoencoder.

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., DeStafano, A. L., Bis, J. C., Beecham, G. W., Grenier-Boley, B., Russo, G., Thorton-Wells, T. A., Jones, N., Smith, A. V., Chouraki, V., Thomas, C., Ikram, M. A., Zelenika, D., Vardarajan, B. N., and et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nat Genet*, 45(12):1452–1458.

Liu, H., Simonyan, K., and Yang, Y. (2019). Darts: Differentiable architecture search.

Mahmoodabadi, S., Ahmadian, A., Abolhasani, M., Eslami, M., and Bidgoli, J. (2005). Ecg feature extraction based on multiresolution wavelet transform. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 3902–3905.

Mayo, C. D., Garcia-Barrera, M. A., Mazerolle, E. L., Ritchie, L. J., Fisk, J. D., and Gawryluk, J. R. (2018). Relationship between dti metrics and cognitive function in alzheimer's disease. *Front Aging Neurosci*, 10:436.

Miranda, M. F., Zhu, H., and Ibrahim, J. G. (2015). Tprm: Tensor partition regression models with applications in imaging biomarker detection.

Mo, C., Ye, Z., Ke, H., Lu, T., Canida, T., Liu, S., Wu, Q., Zhao, Z., Ma, Y., Hong, L., Kochunov, P., Ma, T., and Chen, S. (2022). A new mendelian randomization method to estimate causal effects of multivariable brain imaging exposures. *Pac Symp Biocomput.*, 27:73–84.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. 71(5):1565–1578.

Oh, K., Chung, Y.-C., Kim, K. W., Kim, W.-S., and Oh, I.-S. (2019). Classification and visualization of alzheimer's disease using volumetric convolutional neural network and transfer learning. *Scientific Reports*, 9(1):18150.

Patel, K., Xie, Z., Yuan, H., Islam, S. M. S., Zhang, W., Gottlieb, A., Chen, H., Giancardo, L., Knaack, A., Fletcher, E., Fornage, M., Ji, S., and Zhi, D. (2022). New phenotype discovery method by unsupervised deep representation learning empowers genetic association studies of brain imaging. *medRxiv*.

Raghunath, S., Pfeifer, J. M., Ulloa-Cerna, A. E., Nemani, A., Carbonati, T., Jing, L., vanMaanen, D. P., McCarty, B. E., Hartzel, D. N., Ruhl, J. A., Stoudt, N. J., Johnson, K. W., Zimmerman, N., Leader, J. B., Kirchner, H. L., Griessenauer, C., Hafez, A., Good, C. W., Fornwalt, B. K., and Haggerty, C. M. (2020). Deep neural networks can predict incident atrial fibrillation from the 12-lead electrocardiogram and may help prevent associated strokes. *medRxiv*.

Rasmussen, M. U., Kumarathurai, P., Fabricius-Bjerre, A., Larsen, B. S., Domínguez, H., Davidsen, U., Gerds, T. A., Kanters, J. K., and Sajadieh, A. (2020). P-wave indices as predictors of atrial fibrillation. *Ann Noninvasive Electrocardiol*, 25(5):e12751.

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P. S., Andersson, C. R., Macfarlane, P. W., Meira Jr., W., Schön, T. B., and Ribeiro, A. L. P. (2020). Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature Communications*, 11(1):1760.

Rizwan, M., Whitaker, B. M., and Anderson, D. V. (2018). Af detection from ecg

recordings using feature selection, sparse coding, and ensemble learning. *Physiological Measurement*, 39(12):124007.

Scheff, S. W., Price, D. A., Schmitt, F. A., Scheff, M. A., and Mufson, E. J. (2011). Synaptic loss in the inferior temporal gyrus in mild cognitive impairment and alzheimer's disease. *J Alzheimers Dis*, 24(3):547–557.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Setti, S. E., Hunsberger, H. C., and Reed, M. N. (2017). Alterations in hippocampal activity and alzheimer's disease. *Transl Issues Psychol Sci*, 3(4):348–356.

Shen, L., Kim, S., Risacher, S., Nho, K., Swaminathan, S., West, J., Foroud, T., Pankratz, N., Moore, J., Sloan, C., and et al. (2010). Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in mci and ad: a study of the adni cohort. *NeuroImage*, 53:1051–1063.

Shen, L., Thompson, P., Potkin, S., Bertram, L., Farrer, L., Foroud, T., and for the Alzheimer's Disease Neuroimaging Initiative (2014). Genetic analysis of quantitative phenotypes in ad and mci: imaging, cognition and biomarkers. *Brain Imaging Behav*, 8:183–207.

Sherva, R., Tripodis, Y., Bennett, D., Chibnik, L., Crane, P., de Jager, P., Farrer, L., Saykin, A., Shulman, J., Naj, A., and et al. (2014). Genome-wide association study of the rate of cognitive decline in alzheimer's disease. *Alzheimer's Dement*, 10:45–52.

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum Brain Mapp*, 17(3):143–155.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):1–10.

Taschler, B., Smith, S. M., and Nichols, T. E. (2022). Causal inference on neuroimaging data with mendelian randomisation. *NeuroImage*, 258:119385.

Terza, J. V., Basu, A., and Rathouz, P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3):531–543.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289.

Xu, L., Chen, Y., Srinivasan, S., de Freitas, N., Doucet, A., and Gretton, A. (2020). Learning deep features in instrumental variable regression.

Yang, H., Xu, H., Li, Q., Jin, Y., Jiang, W., Wang, J., Wu, Y., Li, W., Yang, C., Li, X., Xiao, S., Shi, F., and Wang, T. (2019). Study of brain morphology change in alzheimer's disease and amnestic mild cognitive impairment compared with normal controls. *Gen Psychiatr*, 32(2):e100005.

Young, W. J., Haessler, J., Benjamins, J.-W., Repetto, L., Yao, J., Isaacs, A., Harper, A. R., Ramirez, J., Garnier, S., van Duijvenboden, S., Baldassari, A. R., Concas, M. P., Duong, T., Foco, L., Isaksen, J. L., Mei, H., Noordam, R., Nursyifa, C., Richmond, A., and et al. (2023). Genetic architecture of spatial electrical biomark-

ers for cardiac arrhythmia and relationship with cardiovascular disease. *Nature Communications*, 14(1):1411.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57.

Zhang, Y., Xu, Z., Shen, X., Pan, W., and for Alzheimer's Disease Neuroimaging Initiative (2014). Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage*, 96:309–325.

Zhao, B., Li, T., Yang, Y., Wang, X., Luo, T., Shan, Y., Zhu, Z., Xiong, D., Hauberg, M., Bendl, J., Fullard, J., Roussos, P., Li, Y., Stein, J., and Zhu, H. (2021). Common genetic variation influencing human white matter microstructure. *Science*, 372:eabf3736.

Zhao, B., Luo, T., Li, T., Li, Y., Zhang, J., Shan, Y., Wang, X., Yang, Liuqing Yand Zhou, F., Zhu, Z., Initiative, A. D. N., Genetics, P. I. . N., and Zhu, H. (2019). Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nature Genetics*, 51:1637–1644.

Zhao, Q. and Zhang, L. (2005). Ecg feature extraction and classification using wavelet transform and support vector machines. In *2005 International Conference on Neural Networks and Brain*, volume 2, pages 1089–1092.

Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., and Rakovski, C. (2020). A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):48.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552. PMID: 24791032.

Zhu, H., Li, T., and Zhao, B. (2022). Statistical learning methods for neuroimaging data analysis with applications.

Zoni-Berisso, M., Lercari, F., Carazza, T., and Domenicucci, S. (2014). Epidemiology of atrial fibrillation: European perspective. *Clinical Epidemiology*, 6:213–220. PMID: 24966695.

Appendix A

Chapter 2 Appendix

A.1 Proof of Covariate Adjustment Inference

Based on Stage 2, the model can be written as

$$Y = \left(Z\hat{B}_{\hat{\theta}} + W\hat{A}_{\hat{\theta}}\right)\beta + W\gamma + U\alpha + \epsilon$$

For $\hat{\beta}_C$ and $\hat{\gamma}_C$ (the estimate of β and γ with covariate adjustment), based on $Z^\intercal W = 0$,

$$\begin{aligned}
\begin{bmatrix} \hat{\beta}_C \\ \hat{\gamma}_C \end{bmatrix} &= \left(\begin{bmatrix} \left(Z\hat{B}_{\hat{\theta}} + W\hat{A}_{\hat{\theta}}\right)^\intercal \\ W^\intercal \end{bmatrix} \begin{bmatrix} \left(Z\hat{B}_{\hat{\theta}} + W\hat{A}_{\hat{\theta}}\right) & W \end{bmatrix} \right)^{-1} \begin{bmatrix} \left(Z\hat{B}_{\hat{\theta}} + W\hat{A}_{\hat{\theta}}\right)^\intercal \\ W^\intercal \end{bmatrix} Y \\
&= \left(\begin{bmatrix} \hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z\hat{B}_{\hat{\theta}} + \hat{A}_{\hat{\theta}}^\intercal W^\intercal W\hat{A}_{\hat{\theta}} & \hat{A}_{\hat{\theta}}^\intercal W^\intercal W \\ W^\intercal W\hat{A}_{\hat{\theta}} & W^\intercal W \end{bmatrix} \right)^{-1} \begin{bmatrix} \left(Z\hat{B}_{\hat{\theta}} + W\hat{A}_{\hat{\theta}}\right)^\intercal Y \\ W^\intercal Y \end{bmatrix} \\
&= \left(\begin{bmatrix} \left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z\hat{B}_{\hat{\theta}}\right)^{-1} & -\left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z\hat{B}_{\hat{\theta}}\right)^{-1}\hat{A}_{\hat{\theta}}^\intercal \\ -\hat{A}_{\hat{\theta}}\left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z\hat{B}_{\hat{\theta}}\right)^{-1} & (W^\intercal W)^{-1} + \hat{A}_{\hat{\theta}}\left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z\hat{B}_{\hat{\theta}}\right)^{-1}\hat{A}_{\hat{\theta}}^\intercal \end{bmatrix} \right) \begin{bmatrix} \left(Z\hat{B}_{\hat{\theta}} + W\hat{A}_{\hat{\theta}}\right)^\intercal Y \\ W^\intercal Y \end{bmatrix} \\
&= \begin{bmatrix} \left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z\hat{B}_{\hat{\theta}}\right)^{-1}\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Y \\ (W^\intercal W)^{-1}W^\intercal Y - \hat{A}_{\hat{\theta}}\left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z\hat{B}_{\hat{\theta}}\right)^{-1}\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Y \end{bmatrix} = \begin{bmatrix} \left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z\hat{B}_{\hat{\theta}}\right)^{-1}\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Y \\ (W^\intercal W)^{-1}W^\intercal Y - \hat{A}_{\hat{\theta}}\hat{\beta}_C \end{bmatrix}
\end{aligned}$$

proves that $\hat{\beta}_C = \left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z \hat{B}_{\hat{\theta}}\right)^{-1} \hat{B}_{\hat{\theta}}^\intercal Z^\intercal Y = \hat{\beta}_S$, in which $\hat{\beta}_S$ is the estimate of β without covariate adjustment. Then the estimate of σ^2 is

$$\begin{aligned}\hat{\sigma}_C^2 &= \frac{1}{n_{te} - q - w}\left(Y^\intercal Y - \hat{\beta}_C^\intercal \hat{B}_{\hat{\theta}}^\intercal Z^\intercal Y - \hat{\beta}_C^\intercal \hat{A}_{\hat{\theta}}^\intercal W^\intercal Y - \hat{\gamma}_C^\intercal W^\intercal Y\right) \\ &\leq \frac{1}{n_{te} - q - w}\left(Y^\intercal Y - \hat{\beta}_S^\intercal \hat{B}_{\hat{\theta}}^\intercal Z^\intercal Y\right)\end{aligned}$$

The inequality comes from

$$\begin{aligned}\hat{\beta}_C^\intercal \hat{A}_{\hat{\theta}}^\intercal W^\intercal Y + \hat{\gamma}_C^\intercal W^\intercal Y &= Y^\intercal W \hat{A}_{\hat{\theta}} \hat{\beta} + Y^\intercal W \left(\left(W^\intercal W\right)^{-1} W^\intercal Y - \hat{A}_{\hat{\theta}} \hat{\beta}_C\right) \\ &= Y^\intercal W \left(W^\intercal W\right)^{-1} W^\intercal Y \geq 0\end{aligned}$$

Then based on

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\begin{bmatrix}\hat{\beta}_C \\ \hat{\alpha}_C\end{bmatrix}\right) &= \hat{\sigma}_C^2 \left(\begin{bmatrix}\left(Z\hat{B}_{\hat{\theta}} + W\hat{A}_{\hat{\theta}}\right)^\intercal \\ W^\intercal\end{bmatrix}\begin{bmatrix}\left(Z\hat{B}_{\hat{\theta}} + W\hat{A}_{\hat{\theta}}\right) & W\end{bmatrix}\right)^{-1} \\ &= \hat{\sigma}_C^2 \begin{bmatrix}\left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z \hat{B}_{\hat{\theta}}\right)^{-1} & -\left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z \hat{B}_{\hat{\theta}}\right)^{-1} \hat{A}_{\hat{\theta}}^\intercal \\ -\hat{A}_{\hat{\theta}}\left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z \hat{B}_{\hat{\theta}}\right)^{-1} & \left(W^\intercal W\right)^{-1} + \hat{A}_{\hat{\theta}}\left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z \hat{B}_{\hat{\theta}}\right)^{-1} \hat{A}_{\hat{\theta}}^\intercal\end{bmatrix},\end{aligned}$$

we have $\widehat{\mathrm{Var}}\left(\hat{\beta}_C\right) = \hat{\sigma}_C^2 \left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z \hat{B}_{\hat{\theta}}\right)^{-1} \leq \hat{\sigma}_S^2 \left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z \hat{B}_{\hat{\theta}}\right)^{-1}$(neglecting w as $w << n_{te}$) . Thus, the Wald statistic using $\hat{\sigma}_S^2$ as the estimate covariance would be lower than the statistic using $\hat{\sigma}_C^2$, which completes the proof.

Appendix B

Chapter 3 Appendix

B.1 Comparison of Covariate Characteristics

For age (at visit instance 0) and proportion of male individuals, there are differences between the individuals used for GWAS summary statistics computation (without ECG data) and those for DeepFEIVR(-RI) application (with ECG data). For handedness, the two populations are closer. The age difference may come from the fact that the ECG data was collected in visit instance 2 and younger individuals were more likely to participate in the subsequent study.

	GWAS	Training/val/test
Age (mean)	57.10	54.92
Gender (Prop. of male)	45.77%	49.40%
Right-handed (Prop.)	88.69%	88.74%
Left-handed (Prop.)	9.65%	9.71%
Both-handed (Prop.)	1.65%	1.54%

Table B.1: Characteristics of covariates (age, gender and handedness) in the data used for computing GWAS summary statistics and implementing DeepFEIVR or DeepFEIVR-RI.