UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS

GRADO EN BIOTECNOLOGÍA

DEPARTAMENTO DE BASES DE DATOS

Métodos de eliminación de sesgos por covariables en Deep Learning aplicado al campo de las imágenes médicas

TRABAJO FIN DE GRADO

Autor: Gonzalo Cardenal Antolin

Tutores: Ernestina Menasalvas Ruiz, Kerstin Ritter y Roshan Rane

UNIVERSIDAD POLITÉCNICA DE MADRID

Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas

GRADO EN BIOTECNOLOGÍA

CONFOUND DEBIASING METHODS IN DEEP LEARNING APPLIED TO MEDICAL IMAGING

METODOS DE ELIMINACION DE SESGOS POR COVARIABLES EN DEEP LEARNING APLICADO AL CAMPO DE LAS IMAGENES MEDICAS

TRABAJO FIN DE GRADO

Gonzalo Cardenal Antolin

MADRID, 2023

Tutores: Ernestina Menasalvas Ruiz, Departamento de Lenguajes y Sistemas Informáticos e Ingeniería del Software, Universidad Politécnica de Madrid & Roshan Rane, Department of Psychiatry and Neurosciences,Charité - Universitätsmedizin Berlin.

TÍTULO DEL TFG - Métodos de eliminación de sesgos por covariables en Deep Learning aplicado al campo de las imágenes médicas.

Memoria presentada por Gonzalo Cardenal Antolin para la obtención del título de Graduado en Biotecnología por la Universidad Politécnica de Madrid

Fdo: Gonzalo Cardenal Antolin

VºBº Cotutor y Codirector del TFG
Ernestina Menasalvas Ruiz
Dpto. de Lenguajes y Sistemas Informáticos e Ingeniería de Software
ETSIINF - Universidad Politécnica de Madrid

VºBº Cotutores y Codirectores del TFG
Kerstin Ritter & Roshan Rane
Dpto. of Psychiatry and Neurosciences
Charité - Universitätsmedizin Berlin

Madrid, 6, Febrero, 2023

Contents

List of Figures

List of Tables

List of Algorithms

List of Symbols

β	PMDN Layer parameters
$\mathcal{L}$	Lagrangian
$\wedge$	Logical and
$\neg$	Logical negation
$\perp\!\!\!\perp$	Conditional independent
Q	Condtional distribution

List of Abbreviations

ML	Machine Learning
DL	Deep Learning
MVPA	Multivariate pattern analysis
CNN	Convolutional Neural Network
MRI	Magnetic Resonance Imaging
PMDN	Penalty approach for MetaData Normalisation
MDN	MetaData Normalisation
GLM	Generalized Linear Model
UKB	UK Biobank
TM test	Trail Making test
FC	Fully Connected
R^2	Coefficient of Determination
AI	Artificial intelligence
GAN	Generative adversarial networks

SUMMARY

Machine learning algorithms have demonstrated promising results in evaluating clinical psychiatric disorders, and more recently, more advanced deep learning approaches have joined the field. However, these models have been shown to exhibit some biases in their performance. In order to integrate these algorithms into clinical practice, it is important to ensure that they provide robust and unbiased results. This study aims to address the issue of confounding effects in the prediction of various brain phenotypes by applying two debiasing techniques: reweighing and the latest PMDN layer. Our findings indicate that while the reweighing approach was not successful in reducing the impact of confounding effects, the Penalty approach for MetaData Normalisation (PMDN) layer can mitigate the correlation between these confounders and the model's predictions.

RESUMEN

Los algoritmos de aprendizaje automático han demostrado resultados prometedores en la evaluación de trastornos psiquiátricos, y más recientemente, se han sumado aplicaciones de aprendizaje profundo más avanzadas al campo. Sin embargo, se ha demostrado que estos modelos exhiben ciertos sesgos en su rendimiento. Con el fin de integrar estos algoritmos en la práctica clínica, es importante garantizar que brinden resultados robustos y imparciales. Este estudio tiene como objetivo abordar el problema de los efectos de confusión en la predicción de varios fenotipos cerebrales aplicando dos técnicas de eliminación de sesgos: el reequilibrio y la capa PMDN, recientemente desarollada. Nuestros hallazgos indican que mientras que el enfoque de reequilibrio no tuvo éxito en reducir el impacto de los efectos de confusión, la capa PMDN puede mitigar la correlación entre estos confundidores y las predicciones del modelo.

CHAPTER 1. INTRODUCTION AND GOALS

1.1 Introduction

During the last few years, predictive modeling methods such as Machine Learning (ML) and Deep Learning (DL) have made groundbreaking progress in the Medical Image Field. These models have shown to be helpful in the automation of medical image analysis [1],[2]. Especially in clinical Psychiatry, psychiatric disorders enclose high complexity with patient's profile often involving heterogeneous symptomatology [3]. To better describe and clinically assess the complexities of psychiatric disorders, the application of machine learning and other pattern recognition approaches have become increasingly used to harness the rich information observed with human neuroimaging.

Neuroimaging studies of psychiatric and neurological patients used to rely on mass-univariate analytical techniques. These studies typically compared patients with a diagnosis of interest against disease-free individuals and reported neuroanatomical or neurofunctional differences at group level. In the past decade, multivariate pattern analysis (MVPA) has emerged as a popular alternative to traditional univariate analyses of neuroimaging data [4]. The defining feature of MVPA is that it considers patterns of brain activation instead of single units of activation, i.e., for our specific endeavor, voxels in brain Magnetic Resonance Imaging (MRI). One of the most often used types of MVPA is "decoding", in which machine learning algorithms are applied to neuroimaging data to predict a particular stimulus, task, or psychometric feature. Machine learning capitalizes on multivariate data, detecting complex patterns in the brain that may identify abnormalities present in psychiatric disorders [5]. Thus, a wide range of promising medical applications such as diagnosing neurological diseases [6], psychiatric disorders [7], predicting treatment outcomes [8], distinguish disease subtypes [9], or even decoding cognitive and behavioral factors [10] have emerged.

Across the many different ML methods, a common distinction is drawn between two types: supervised and unsupervised learning [11]. In supervised learning, a ML algorithm is trained to assign a probability of a data point belonging to a certain category (labeled as "Y") based on its features (labeled as "X"). After training, the ML algorithm can correctly assign these labels to new data. In contrast, unsupervised learning does not require labeled training data.

Instead, it can use unlabeled data, such as whole-genome sequences or unlabeled MRI scans, to find clusters within these data points. For example, from the variety of aforementioned tasks, the diagnosis of a neurological disease falls under the category of supervised learning, whereas the identification of disease subtypes is classified as an unsupervised learning strategy.

Major breakthroughs in the field of natural image recognition have also provided specialized deep learning architectures to successfully study these neuroimages [1]. One example of such an architecture is Convolutional Neural Networks (CNNs). Recently, CNNs set a baseline of solid architectures for prediction tasks in Computer Vision. The main novelty CNNs introduced was convolution. Mathematically, a convolution is an integration function that expresses the amount of overlap of one function "g" as it is shifted over another function "f". In a Neural Network, this operation is implemented through the use of "kernels", which are small matrices of weights. These kernels are applied to the input data by sliding over it and producing a weighted sum as the output. This weighted sum, also known as the feature space, is then used as input for subsequent layers in the network. CNNs are able to extract low-dimensional features from input data by adjusting the weights in these fixed-size kernels.

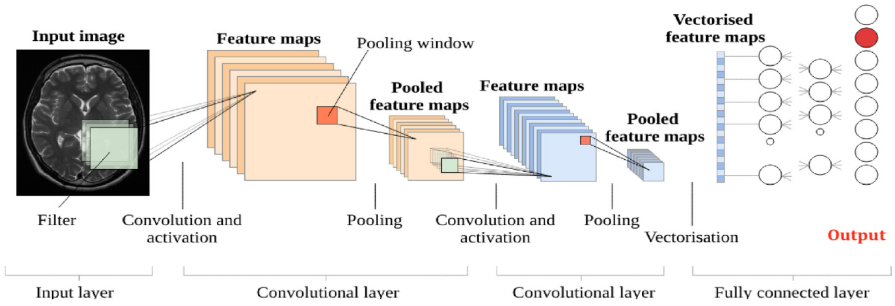

Figure 1: CNNs architecture with MRI as input sample. Model flow is depict, kernels filter the image to obtain feature maps and output a vector. Slight modification from [12].

Nevertheless, while ML & DL hold promise as tools for evaluating psychiatric disorders, these approaches also come with unique challenges and trade-offs. One of these challenges is addressing confounding effects[13]. A fundamental limitation of decoding analyses is that it remains ambiguous which source of information drives decoding performance, which becomes problematic when the to-be-decoded variable is confounded by variables that are not of primary interest [5]. In statistics, a confounder is a variable that influences both the dependent

and independent variables, causing a spurious association. In ML with respect to predictive neuroimaging models, a confounding variable c is defined as a variable that correlates with the target y and is deducible from the input X, and this relationship $X \rightarrow c \rightarrow y$ is not of primary interest to the research question and hinders the analysis [14] (see Fig.2). Besides, in the framework of Fairness in medical Artificial Intelligence (AI), these confounders can also be understood as sensitive attributes [15]. In a simple classification task, the patient's information can be separated as task-related information (medical images, denoted as X), e.g., MRI images, and task-irrelevant information that is inherent, such as age, sex, race, etc. This irrelevant information is called sensitive attributes. In certain cases, certain sensitive attributes may be related to the target task, in which case, the concept of confounders and sensitive attributes converge. Hence, these confounders or sensitive attributes can influence the performance of the model predictions. It is crucial to note that even if certain sensitive attributes may be related to the target task in some cases, they should not be utilized for categorization in the classification task as this can lead to algorithms relying on the easiest criterion for classifying samples, a phenomenon known as shortcut learning. Shortcut learning can occur as a result of inadequately addressing confounding factors. This occurrence can also result in unfair models.

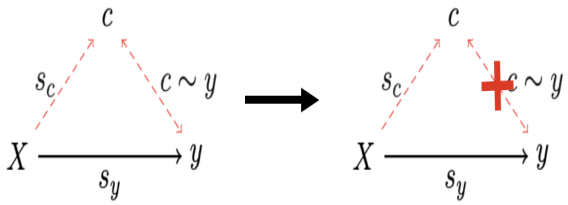

Figure 2: Confounders Causal Graph. Visualization of the correlation between confounding variables and the relationship that needs to be removed.

To illustrate this issue, let's consider the prediction of alcohol misuse. Greater proportion of males are risky alcohol users compared to females [14]. These systematic differences can interfere with ML analyses because ML models may use the sex information present in neuroimaging data to indirectly predict alcohol abuse rather than directly identifying the effects of alcohol on brain structure. As a result, our model is biased and unfair. For instance, fluid intelligence is strongly impacted by age; age is well predicted from brain images; hence successful prediction of fluid intelligence from brain images might captured nothing more than a biomarker of aging [16]. Another example is models that predict Alzheimer's disease

from brain MRI; these models were found to be confounded by measurement artifacts such as scanner strength [17], and the demographics of the subjects such as age and sex [18].

In ML, common confound debiasing methods like regressing out the confounding signal from the input or post hoc counterbalancing the data are popular, effective strategies. However, in DL models, the problem not only gets exacerbated due to the non-linearity in the model [19] and the stochasticity in a DL optimization process but also there is a lack of standardized strategies to address it. Some incipient research shed light on this issue. Two promising methods have been recently implemented for DL: reweighing [20] and PMDN [21].

This work evaluates within the framework of predictive modeling in neuroimaging the performance of these two confound debiasing techniques for DL models over different confounded tasks. These models were trained on structural MRI data and aimed to predict different brain phenotypes.

1.2 Goal & Objectives

The main goal of this work is to successfully implement two confound debiasing methods in a CNN training pipeline. To fulfill this goal, the following objectives are achieved:

1.2.1 Identification of Confounding Variables in the prediction of the selected brain phenotypes

To eliminate any confounding signals from the model's features, it is necessary to first demonstrate the presence of such signals in the data and their impact on the predictions.

1.2.2 Application of our methods in the training to obtain unbiased models

The models are retrained using the same heuristics, but incorporating the debiasing methods. Mantaining heuristics allows for systematic comparisions.

1.2.3 Evaluation of methods performance to the unbiased confounded models

To confirm the success of the implementation, the performance of the methods debiasing the confounding variables is assessed.

CHAPTER 2. MATERIALS AND METHODS

2.1 Experiments Pipeline

As depicted in the diagram 3, this thesis comprises of two distinct stages: "Confounders Exploration" and "Debiasing Techniques Application." The former focuses on identifying potential confounding factors that may impact the target phenotypes, the latter addresses the issue by utilizing the selected methods: "Reweighing" and "PMDN".

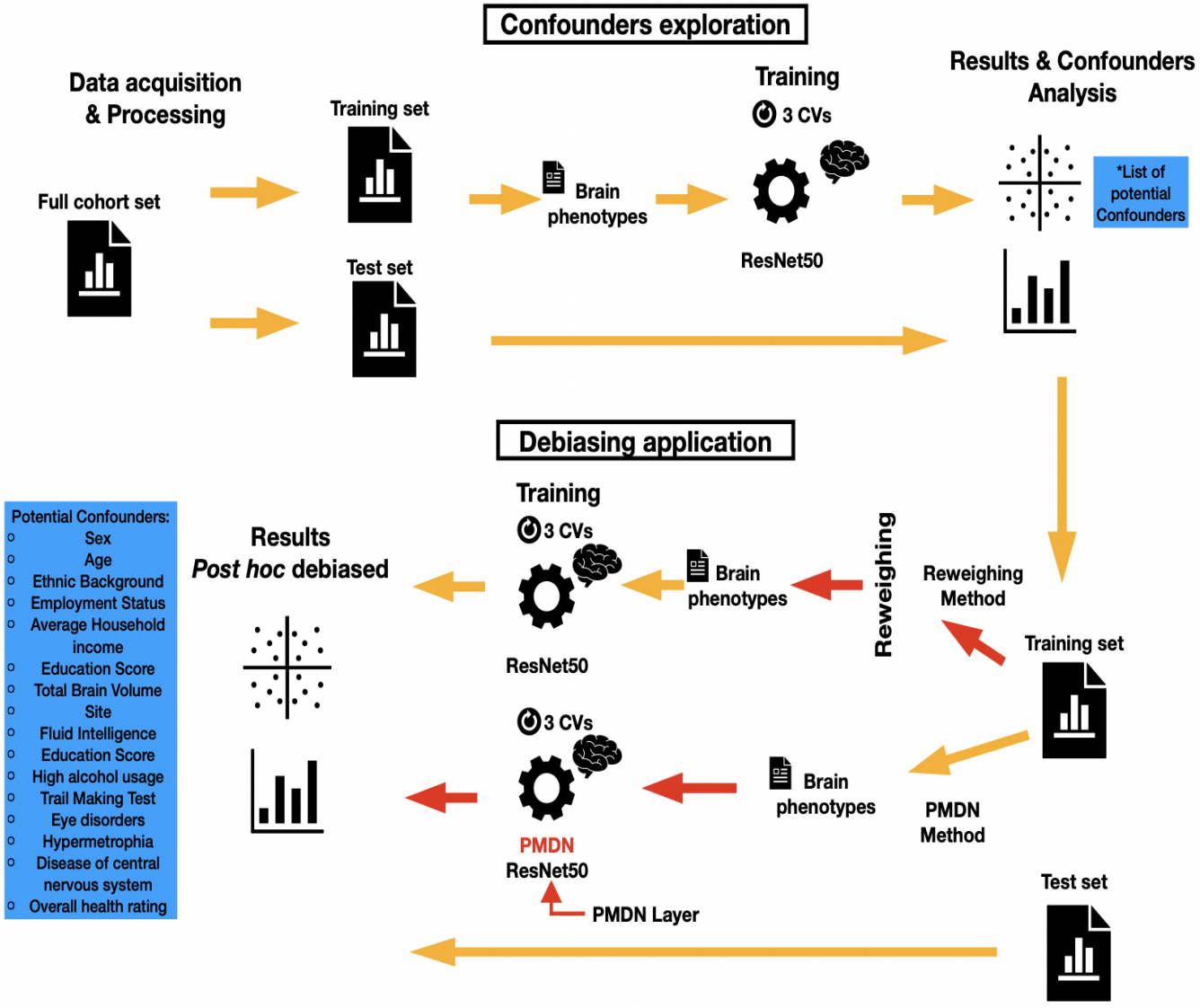

Figure 3: Project Pipeline. The study is divided in two stages: Confounders exploration and Debiasing application. Red arrows indicate the steps where debiasing methods where implemented. The blue list illustrates the variables studied as potential confounders.

In **"Confounders Exploration"**, we divide the subjects into a training and test set. Using the training set, we train a CNN network to predict various brain phenotypes, including sex, age, high alcohol usage, and Trail Making test. The trained model is then applied to the test set to generate predictions. We calculate scores to quantify the correlation between the potential confounders and the predictions.

The stage of **"Debiasing Techniques Application"** involves repeating the previous training upon implementation of the debiasing strategies. Once the primary confounding variable for each predictive task is identified, we apply the chosen methods to eliminate the encoded information of this specific confounding variable whose signal is biasing the training process. **"Reweighing"** is a pre-processing method that is applied to the training set to compute a set of weights that will correct for the loss values during the training due to the difference in expected and the observed probability of a given phenotype. **"PMDN Layer"** is an in-processing method that operates as an integrated layer in the network architecture, filtering the confounders signal on the feature representations. A more detailed explanation and formal definition of these methods can be found in Section 2.5. Lastly, after applying the predictive trained model to the test set, the $post - hoc$ debiased results are analyzed to evaluate the effectiveness of the two methods.

2.2 Dataset and Image Processing

The first stage of the thesis, which relied on confounders exploration, required a wide set of additional information, defined as metadata or labels (demographics, cognitive, behavioral, acquisition variation, etc.). Hence, we based our analyses on the UK Biobank (UKB)[22]. The UKB is a large biomedical database deemed the "world's largest multi-modal imaging study" [23] and offers genotyping and phenotyping data on approximately half a million participants, with 46197 (as of the June 2021 release) having undergone additional medical imaging. This additional medical imaging collection was conducted in three assessment centers to ensure imaging homogeneity and minimize confounding effects due to the location of the assessment centers; identical scanner hardware, and software were used in all sites [24]. For further information regarding data acquisition and processing protocols, please refer to [25].

To further reduce the potential confounding effects of site-specific acquisition variables, we

divided our dataset into a training set and a test set using the UKB assessment center as the criterion. The training set consisted of all subjects from Cheadle and Reading, totaling 30,593, while the test set included the whole cohort from Newcastle, comprising 10,089 subjects. We then defined a list of potential confounders from the available metadata and excluded all subjects for which this information was unavailable. This resulted in a training set of 8,364 subjects and a test set of 3,460 subjects, each containing structural T1-weighted brain images of resolution size 182x218x182. As a further processing step, the images were downsampled to 52.5% of their initial resolution, resulting in a sample size of 96x114x96 4.

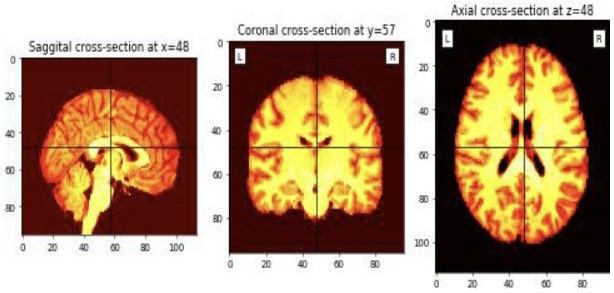

Figure 4: T1-weighted MRI Data sample.3D slice from subject 38 with resolution 96x114x96.

2.3 Brain phenotypes and Potential Confounders

Four tasks were selected as representative variables to study their confounders: Sex, Age, Trail Making Test (TM Test) [26], and High Alcohol Usage. Sex and High Alcohol Usage are binary data types and, therefore, are a classification task, whereas Age and TM Test are continuous data types that necessitate a regression analysis.

Previous research has demonstrated the ability to predict sex and age from T1-weighted MRI [27]. Brain size is a potential confounding factor for sex prediction, as a high correlation has been established between brain size and sex, with males generally having larger total brain volume than females [28]. On the other hand, age's primary potential confounding factor is also brain volume, as aging has been shown to be associated with a decrease in this attribute [29]. The Trail Making Test (TM Test) is a widely used neuropsychological instrument, which can be employed as a standalone tool for detecting neurological disease and neuropsychological impairment [30]. It assesses cognitive domains such as processing speed, sequencing, mental flexibility, and visual-motor skills. Therefore, we anticipate that it may be confounded with

factors like age or vision-related diseases, such as hypermetropia or other eye disorders. High alcohol usage is assessed through UKB questionnaires, which provide information on the amount and frequency of alcohol consumption. A higher proportion of males tend to exhibit high-frequency drinking behavior, so we expect this to be confounded with sex [14].

In order to conduct a comprehensive examination of confounding variables, we compiled a list of potential confounding factors from various sources, including demographic information, brain characteristics, data acquisition, health, and cognitive factors. Table 1 presents a summary of the confounding variables studied.

Table 1: Potential Confounders

Confounder	Potential confounded target	UKB ID
Demographics		
Sex	Age, High alcohol usage	31
Age	TM Test	21003
Ethnic Background	High alcohol usage, sex	21000
Employment Status	Age	6142
Average Household Income	Age	738
Education Score	Age, TM Test	26414
Brain attributes		
Total Brain Volume	Age, sex	26521
Data acquisition		
Site	All	54
Behavioural		
Handedness	All	1707
Tobacco Smoking	Age, High alcohol usage	1239
High Alcohol Usage[a]	Sex	1558,20414,20403,20416
Health		
Eye disorders	TM Test	6148
Hypermetrophia	TM Test	5832
Disease of central nervous system[b]	Age	ICD10 - 41270
Overall health rating	Age, High alcohol usage	2178
Cognitive		
Fluid intelligence	Age	20016
Trail Making test[c]	Age	6348 & 6350

[a] Bahviour determined with the set of IDs
[b] CNS disease interval: [G00-G47][G58-G64][G80-G99]
[c] Data used is their mean value

2.4 Network architecture and Training

In the experiments, we use the ResNet50. Residual Networks or ResNets, are well-known standardized architectures in the framework of CNNs. By using residual connections between convolutional blocks these networks address the problem of the vanishing gradient and overfitting. ResNets have numerous variants that vary in the number of layers, ReNet50 comprises

50 neural network layers. The model's architecture is depicted in Figure 5. For additional information, please refer to [31].

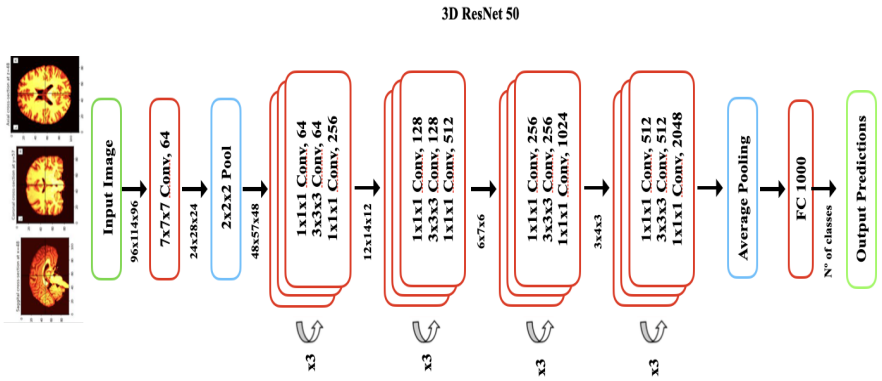

Figure 5: ResNet50 Architecture. An FC layer refers to Fully Connected layer. Model's output shape corresponds to N, where N is the number of classes.

The network was trained using the Adam optimizer [32] with a learning rate of 5×10^{-4} and weight decay of 10^{-4}. We decayed the learning rate with a scheduler of step size 20 and gamma 0.1. The batch size was set to 24, and in order to facilitate faster convergence, pretrained weights were loaded from a model trained on videos[33]. During training, brain voxel values were normalized and the data was augmented by flipping the images across the sagittal plane and translating along the sagittal axis by up to two voxels in either direction, with both methods being standard data augmentation methods for deep learning methods performing medical image analysis [34],[35].

We performed 3-Fold Cross-Validation [36] where the maximum number of epochs for each fold was 120; nevertheless, training was stopped once the accuracy metric achieved by the model on the validation set did not improve over fifteen epochs, after which the model state with the highest validation accuracy metric was evaluated on the test set. To enhance the reproducibility of the results, a random seed number of 46 was utilized.

The selection of the loss function and validation metric is contingent on the specific task at hand. In the case of binary classification, we use Binary Cross Entropy with Logits loss, and the prediction performance is estimated using Balanced Accuracy, which accounts for class imbalance. For regression tasks, the Mean Squared Error loss and R^2, which measures the

explained variance, are estimated instead.

2.5 Confound debiasing methods

According to [15] and [20], DL pipelines contain three possible points of intervention to mitigate unwanted bias: the training data, the learning procedure, and the output predictions, and these are associated with three corresponding classes of bias mitigation strategy: pre-processing, in-processing, and post-processing.

Pre-processing methods focus on the quality of the training dataset; they modify the training dataset to remove the correlation before training the AI model. Common strategies used are under-sampling and over-sampling [37], stratified batch [20], or use Generative Adversarial Networks (GANs)[38] to generate additional images to correct for imbalanced datasets. Additionally, **reweighing**, also referred to as sample weighing, is another effective strategy, and it is the one considered in this study.

In-processing methods aim to modify state-of-the-art learning algorithms in order to remove confounders signal during the model training process via model regularization. In this direction, most of the confound debiasing approaches have been proposed [39], [40], [41]. Among these, the **PMDN Layer** is one of the latest and most promising and is evaluated in this work.

Lastly, **post-processing** approaches [42],[43] correct the output of an existing algorithm to satisfy the confounder-free requirements. In this study, no post-processing method was utilized as the goal was to obtain confounder-free representations within the model.

2.5.1 Reweighing

The method of reweighing was initially introduced in [44] as a pre-processing technique to remove discrimination. This method was then adapted in the AIF360 Toolkit [45]. However, in both settings, the implementation of the method is limited to binary confounders. In this work, we extend it to include multiclass confounders, thereby increasing its applicability.

In this method, we assign weights to specific labels. For a given dataset D with N samples $d_1, d_2, ..., d_N$, the i-th sample d_i consists of medical image data $\mathbf{X_i}$, a target task ground

truth label $\mathbf{Y_i}$, and a list of metadata variables which we denotate as possible confounders $C = C_1, C_2, ..., C_L$ containing L elements, i.e. $d_i = (X_i, C_i, Y_i)$. Due to algorithms limitation, we discretise the continuous confounders. We denote the number of possible values in $\mathbf{Y_i} \in (T_1, T_2, ..., T^N)$ being the number of values in the truth label, and $\mathbf{C_i} \in (G_1, G_2, ..., G^T)$ as the number of possible values by the confounder. Thus, the idea underlying weight calculation is the following one:

Having an unbiased D, i.e., $\mathbf{Y_i}$ and $\mathbf{C_i}$ are statistically independent, the expected probability $Pexp(\mathbf{Y_i} = T_n \wedge \mathbf{C_i} = G_t)$ would be:

$$Pexp(\mathbf{Y_i} = T_n \wedge \mathbf{C_i} = G_t) := \frac{|X \in D | X(\mathbf{Y_i}) = T_n|}{|D|} \times \frac{|X \in D | X(\mathbf{C_i}) = G_t|}{|D|} \tag{1}$$

In reality, however, the observed probability in D,

$$Pobs(\mathbf{Y_i} = T_n \wedge \mathbf{C_i} = G_t) := \frac{|X \in D | X(\mathbf{Y_i}) = T_n \wedge X(\mathbf{C_i}) = G_t|}{|D|} \tag{2}$$

may deviate from the expected. If the expected probability exceeds the observed probability value, it indicates a bias towards $X(\mathbf{C_i}) = \neg G_t$ for those instances of X for which $X(\mathbf{Y_i}) = T_n$.

To compensate for the bias, we will assign lower weights to objects that have been deprived or favored. Every object X will be assigned weight:

$$\mathbf{W}(\mathbf{X_i}, \mathbf{C_i}, \mathbf{Y_i}) := \frac{Pexp(\mathbf{Y_i} = T_n \wedge \mathbf{C_i} = G_t)}{Pobs(\mathbf{Y_i} = T_n \wedge \mathbf{C_i} = G_t)} \tag{3}$$

i.e., the weight of an object will be the expected probability to see an instance with its confounder value and true label given independence, divided by its observed probability. In this way we assign a weight to every tuple according to its $\mathbf{Y_i}$ and $\mathbf{C_i}$-values. This results in a new dataset, $\mathbf{D_w}$, in which confounding effects of the given $\mathbf{C_i}$ are eliminated by introducing these computated weights as multiplying factors for each tuple class ($\mathbf{Y_i}$,$\mathbf{C_i}$). A $\mathbf{C_i}$-confounder-free classifier can then be trained on this balanced dataset.

The algorithm outlining the approach to Reweighing is presented in **Algorithm 1**, as represented in pseudocode.

Algorithm 1: Reweighing

Input: (D, L, C)
Output: Classifier learned on reweighed D

```
for l ∈ Label categories do
    for c ∈ Confounder categories do
        W(l, c) := P_exp(L=l∧C=c) / P_obs(L=l∧C=c)
    end
end
D_w :=
for X in D do
    Add (X, W(X(L), X(C))) to D_w
end
Train Classifier CNN on training set D_w, introducing weights in the loss function
return Classifier CNN
```

As an important concern, this method is limited to categorical variables. As classes for continuous variables can not be counted, which is the main foundation behind the algorithm, when performing reweighing for a regression prediction task the continuous variable had to be binarized, and when correcting for a continuous confounder, the continuous variables had to be discretized in multiple categories. In our experiments, we binarized continuous phenotypes by median, and we discretized continuous confounders using a Decision Tree Discretiser, a supervised discretization method that minimizes information loss during discretization and maximizes the predictive performance of the model[46].

2.5.2 PMDN Layer

The Penalty approach for Metadata Normalization Layer (**PMDN Layer**) [21] is a recently proposed method that builds upon the MetaData Normalization (**MDN**)[47]. The MDN estimates the linear relationship between the metadata and each feature by creating a Generalized Linear Model (GLM) $\mathbf{f} = \mathbf{X}\beta + \mathbf{r}$, where β is a learnable set of linear parameters, $\mathbf{X}\beta$ corresponds to the component of $\mathbf{f}$ explained by the metadata, and r is the residual component irrelevant to the metadata. This strategy is implemented as a non-trainable closed-form solution, meaning that it requires the building of a linear model (relationship between metadata and each feature) on a batch-level operation, where confounding effects are removed from the

features within the training batch.

The main **limitation** of this approach is that the GLM must learn the parameters within the batch. As a result, small batch sizes can cause instability during training, as large batch sizes are required to obtain accurate approximations of the linear estimator. Therefore, the MDN approach is only effective if the training setup of the model has high computing power. In our specific usecase, our computation systems do not allow for big batches, making the MDN approach infeasible. The PMDN is a **promising alternative** to this limitation. To overcome this limitation, PMDN introduces a penalty method that transforms the MDN approach into a layer with parameters that can be optimized with other components of the network during training. This turns PMDN into an open-form solution.

Let $\mathbf{X} = [x_1, x_2, ..., x_N]^T$ be the N training samples and $\mathbf{y} = [y_1, y_2, ..., y_N]^T$ be their corresponding prediction target labels. Without loss of generality, let us assume that a network can be defined as the composition $\psi(\phi(\mathbf{X}))$ between the first few layers ϕ and the layers afterwards ψ. For simplicity, we assume ϕ results in a one-channel feature but the following formulation generalizes to multi-channel features. Let $\mathbf{W}$ be the network parameters of ψ and ϕ, then training of the network often reduces to solving the minimization problem :

$$\min_{W} \mathcal{L}(\psi(\phi(\mathbf{X})), y) \tag{4}$$

When a linear estimator (GLM from the MDN) is introduced after the layers represented by ϕ, the minimization problem changes to:

$$\min_{W} \mathcal{L}(\psi(\phi(\mathbf{X}) - \mathbf{M}\beta_{ls}), y) \tag{5}$$

$$s.t \beta_{ls} = arg\min_{\beta} \mathcal{L}^*(\phi(\mathbf{X}); \mathbf{M}) = arg\min_{\beta} ||\phi(\mathbf{X}) - \mathbf{M}\beta||^2 \tag{6}$$

In simpler terms, the constraint is a type of optimization within an optimization, whose goal is to eliminate the influence of metadata as much as possible from the features learned by ϕ. The PMDN adds a penalty term to this optimization to find a solution to the bi-level optimization problem by determining the minimum value of a proxy objective function that

combines the two minimization problems:

$$\min_{\beta,W} \mathcal{L}(\psi(\phi(\mathbf{X}) - \mathbf{M}\beta_l s), y) + \lambda \mathcal{L}^*(\phi(\mathbf{X}); \mathbf{M}) \tag{7}$$

Equation 7 is a mathematically well-defined and differentiable function that can be optimized using any gradient descent algorithm. The PMDN method allows for the estimation of β to converge to a local optimum that is defined with respect to the entire training data set. According to **Algorithm 2**, both parameters to optimize (line 5 and 9) have their own learning rates, and are then consolidated into the optimizer (in our case Adam). It should be noted that this implementation makes it independent from λ hyperparameter.

In our experiments, one PMDN Layer with β learning rate of 0.02 is implemented using our ResNet50 architecture between the output of the last convolutional block and the Average Pool layer (fig.6). We define this model PMDNResNet50.

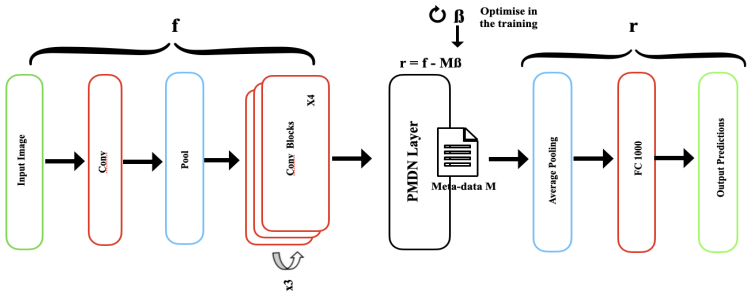

Figure 6: PMDNResNet50 architecture. A PMDN layer is introduced between the last convolutional block and the average pooling layer. The confounders label is incorporated into this layer, and its strength to subtract f is modulated by the β parameters, which are optimized during training.

2.6 Statistical tests to quantify confounding effects

Quantifying confounding bias in predictive models poses a statistically significant challenge due to the presence of non-normality and nonlinearity in the model's output. To identify the possible confounding variables and quantify the scale of their influence in prediction, we use the framework proposed by [48]. This strategy consists of two Statistical Test, the Partial Confounder Test (PCT) and the Total Confounder Test(TCT).

In a predictive modeling setting, where y denotes the target variable, $\hat{y}$ the model output and c the confounder variable, the conditional independence between $\hat{y}$ and c given y is $\hat{y}$ $\hat{y} \perp\!\!\!\perp c|y$,

Algorithm 2: PMDN Layer

Input: $(W, \beta, \eta_1, \eta_2)$

Output: Classifier learned on reweighed D

for *t in (0,1,...,T)* **do**

  Freeze $W^{(t)}$, Unfreeze $\beta(t)$

  $\hat{y}$=$\psi(\phi(\mathrm{X})$ - $\mathrm{M}\beta^t)$ ▷ `Foward pass`

  $\beta^{t+1} = \beta^t - \eta_1 \nabla_{\beta^{(t)}} \mathcal{L}^*(\phi(X); M)$

  Freeze $\beta(t+1)$, Unfreeze $W^{(t)}$

  $\hat{y}$=$\psi(\phi(\mathrm{X})$ - $\mathrm{M}\beta^{t+1})$ ▷ `Foward pass`

  $W^{(t+1)}$= $W^{(t)}$ -$\eta_2 \nabla_{W^{(t)}} \mathcal{L}(y, \hat{y})$

end

return Classifier CNN with PMDN Layer

which, by definition, means that $P(\hat{y}, c|y) = P(\hat{y}|y)P(c|y)$. Determining the independence of c from $\hat{y}$, given y, involves verifying if the route $c \to X \to \hat{y}$ has been interrupted in the prediction algorithm. This statistical evaluation, with the null hypothesis $H0 : \hat{y} \perp\!\!\!\perp c|y$, is referred to as the *Partial Confounder Test*. One might also be interested in testing $\hat{y} \perp\!\!\!\perp y|c$. We refer to the corresponding test as the *Full Confounder Test*.

Within this setting, we are investigating two distinct null hypotheses corresponding to the $(y, \hat{y}, c)$ triplet. Testing the null hypothesis $\hat{y} \perp\!\!\!\perp c|y$ examines whether the dependence of the model output on the confounder can likely be explained by the confounder's dependence on the target variable (i.e., whether there is any confounding bias in the model). Testing the null hypothesis $\hat{y} \perp\!\!\!\perp y|c$ determines if the predictions can be solely explained by the confounder (i.e., whether the model is exclusively confounder driven). However, without assumptions on the joint distribution of $(y, \hat{y}, c)$, conditional independence testing is not effectively impossible. Hence, the PCT requires the assumption$(Q(c|y))$ and the FCT that $(Q(y|c))$.

Table 2: Conditonal Independance Confounder Tests

	H0	**Assumption**	**Output**	**Results Interpretation**
PCT	$\hat{y} \perp\!\!\!\perp c\|y$	$Q(c\|y)$	$R^2_{\hat{y}c}$, p-value	$R^2_{\hat{y}c} > 0.3 \wedge p < 0.01$ = Cofounded model
FCT	$\hat{y} \perp\!\!\!\perp y\|c$	$Q(y\|c)$	$R^2_{\hat{y}y}$, p-value	$p > 0.01$ = Fully confunded model

CHAPTER 3. RESULTS AND DISCUSSION

3.1 Model's performance

After the training, model's performance for the different predictions of brain phenotypes is displayed in Table 3. For binary classification tasks (Sex and High Alcohol Usage), the metric used to measure performance was binary balanced accuracy, whereas for regression tasks (Age and Trail Making Test) was R^2, commonly known as the coefficient of determination. Thus, for binary labels the probability to correctly predicting a class by chance, i.e., chance accuracy, is 0.5 and for regression 0.0. Higher results than chance accuracy imply a significant model's predictive performance.

Table 3: Model's Performance. This table shows the Balanced Accuracy values for binary and Explained Variance for regression of the different prediction tasks.

Phenotype	Validation Set	Test Set
Sex	0.9982	0.9884
Age	0.922	0.7917
High Alcohol Usage	0.5	0.5
High Alcohol Usage *	0.5997	0.5763
Trail Making Test	0.086	0.0002

*values correspond to a second training with a bigger cohort of 14617 samples

Examination of Table 3 shows that the model predicted sex with a balanced accuracy of 0.9982 and 0.9884, and age with a coefficient of determination of 0.922 and 0.7917, for the validation and test sets, respectively. The prediction of sex and age as brain phenotypes has been previously established in the literature [27]. Our results confirm the effectiveness of the ResNet50 in accurately predicting sex. The accuracy of age prediction was high for the validation set, but the model's performance was not as robust for the test set, as demonstrated by the decrease in performance.

In contrast, the prediction of high alcohol usage and the TM test has been attempted before [14], but they present a challenging task. The classification of high alcohol users was unsuccessful, with the model's accuracy being no better than chance. This could be due to insufficient information in the MRI scans of the training cohort, the phenotype being unpredictable for all possible combinations of cohorts, or insufficient sample size to refine the

model's parameters. To address this, a larger training cohort of 14,617 samples was used to retrain the ResNet50 model. The data processing followed similar heuristics to the previous training cohort of 8,617 samples, but a larger sample size was prioritized over maintaining meta-data constraints. As a result, only sex and age were considered potential confounders for high alcohol usage, as other phenotype variables were not available for the entire cohort. This training is denoted in the table by High Alcohol Usage*.

The results showed that neither task was predicted with an accuracy greater than chance by 10% or more. The model performed particularly poorly on the TM test, with an R^2 of 0.02% on the test set, suggesting a complete failure to generalize. Despite this, the model was still able to extract information from the MRI scans to make meaningful predictions. Therefore, we analysed what information drives these predictions by examining their potential confounders.

3.2 Analysis of the potential confounders

In order to assess the possibility of confounding effects, we performed a systematic Partial Correlation Test on the model's output ($\hat{y}$), the current predicted label (y), and the label of the confounder associated with each subject (c). The test yields the R^2_{yc} value, which represents the coefficient of determination between the confounding factor c and the predicted variable $\hat{y}$. This coefficient indicates the extent to which the variation in the dependent variable can be predicted based on the independent variable. Specifically, R^2_{yc} reflects the degree to which the prediction of $\hat{y}$ can be accounted for by the confounding factor c, which is the relationship we aim to eliminate (see fig.2). These values are presented in a bar plot in fig.7). Bars with asterisks indicate a R^2_{yc} with p-value less than 0.001.

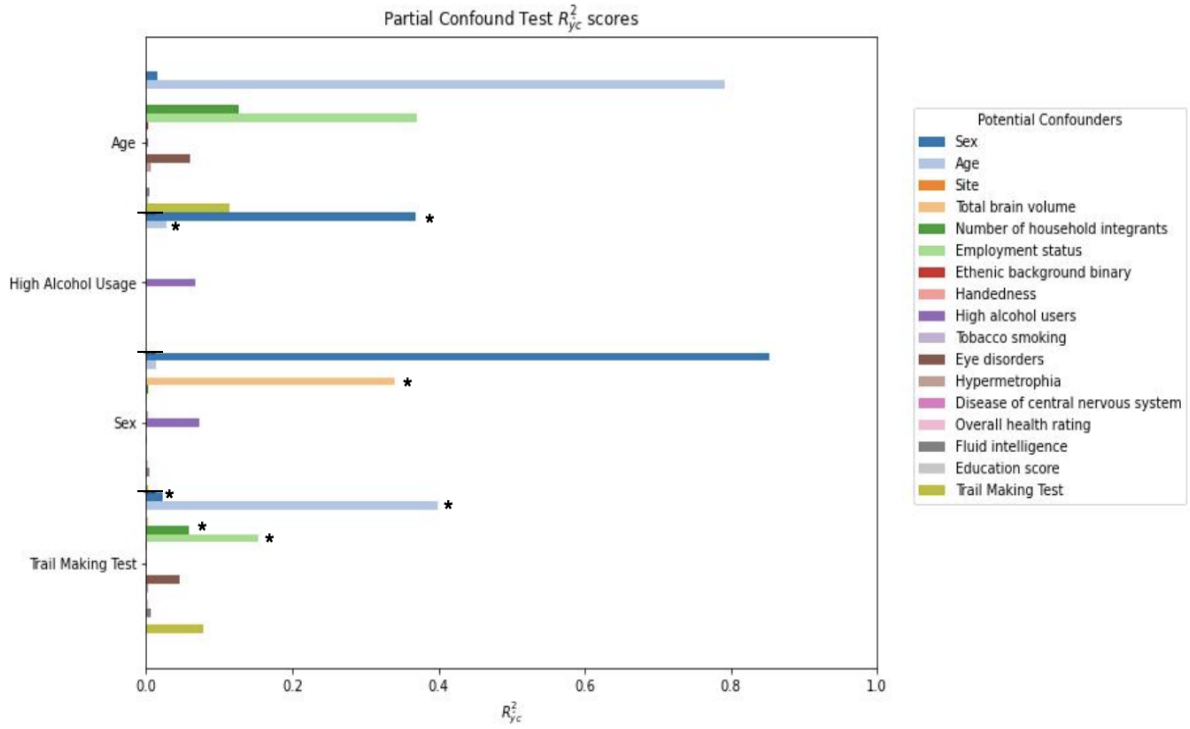

Figure 7: Systematic analysis of potential confounders. The bars display the R^2_{yc} of the PCT between the brain phenotype and the potential confounder variable. * indicates p-value< 0.0001.

Firstly, it is noteworthy that fig.7 also displays the PCT between the predicted label and itself. A high R^2_{yc} and a high p-value on this test indicate that the test is functioning correctly and allows for comparison. High p-values for this test are crucial, as low values would suggest that the phenotype is being confounded by itself, which is impossible and would indicate a malfunction of the PCT. Furthermore, a higher R^2_{yc} in a PCT that is not the one run with the predicted label implies that the performance of the model is heavily influenced by the confounder rather than by the predicted label itself.

In the analysis of the first brain phenotype, age, it was found that the variables with a coefficient of determination greater than 10% were Number of Household Members, Employment Status, and Trail Making Test. This outcome is reasonable, as increased age often results in a shift from employment to retirement and a decrease in performance on the Trail Making Test. The relationship between age and the number of household members in the UKB dataset is likely due to the age range of the subjects recruited (42-82 years old, see Annex B.1 to check data distribution of variables) - in this age range, household size is known to decrease as chil-

dren, or partners leave the household, whether due to growing up or health reasons. However, none of these variables had p-values less than 0.01, and thus, although there is a correlation among the variables, they do not significantly affect the model's prediction of age.

In the case of High Alcohol Usage, the analysis was limited to age and sex. Both variables were found to have p-values less than 0.0001, thereby hindering the classification of high alcohol users. Of the two, sex had the highest R^2_{yc}, even surpassing the predicted label itself. Therefore, it was determined to be the primary confounder.

For Sex, only total brain volume demonstrated a R^2_{yc} greater than 10% and a p-value less than 0.001. Although a correlation with high alcohol usage can be observed, it did not meet the criteria established for total brain volume. Thus, total brain volume was identified as the primary confounder.

With respect to the Trail Making test, all variables correlated with age were deemed significant confounding factors. Age was clearly identified as the primary confounder, as it had a R^2_{yc} of 0.4 and a p-value less than 0.0001.

Prior to conducting the tests, we anticipated that sex would be confounded with total brain volume, high alcohol usage with sex, and the TM test with age. Our hypothesis regarding the main confounders for sex, high alcohol usage, and the TM test was therefore confirmed. Conversely, the results of the comprehensive confounder analysis indicate that age is not confounded and that there is no requirement to account for any of the analyzed confounders.

3.3 Analysis of the primary confounders

In order to determine if the predictions of the trained models are solely influenced by the primary confounders or only partially affected, we perform the Full Confounder Test. This test generates $R^2_{y\hat{y}}$ and corresponding p-value. A high p-value in the relationship y $\hat{y}$ signifies a lack of statistical significance and that the model predictions are completely driven by the confounder. In this study, we arbitrarily set the threshold for full confounding to a p-value greater than 0.1; however, there is no established consensus in the literature on what p-value is optimal for this classification. The results of this test on the primary confounders are shown in Table 4.

Table 4: Full Confound Test. p-value>0.01 indicate low statistical significance between y and $\hat{y}$ which suggest fully confounded model.

Full Confound Test			
Phenotype	Confounder	$R^2_{\hat{y}y}$	p value
Sex	Total brain volume	0.821	0.0001
High Alcohol Usage	Sex	0.067	0.0001
Trail making Test	Age	0.079	0.03

The p-values for sex and high alcohol usage were less than 0.001, which signifies that their coefficients of determination are statistically significant and the predictions are not completely driven by the confounders. The outcome for sex was as expected, as in the PCT the highest $R^2_{\hat{y}c}$ was found to be higher for sex compared to total brain volume. However, in the case of high alcohol usage, the highest $R^2_{\hat{y}c}$ was found to be with sex rather than with itself. This provides further evidence that the model can decode signals from T1-weighted MRIs specific to the alcohol misuse phenotype.

For the TM test, the p-value exceeded the threshold, indicating that the model only captured the signals of the confounders. Moreover, as displayed in Table 3, the explained variance on the test set was not statistically significant, suggesting that the relationship between the TM test and age that was present in the training data may no longer exist in the test data, resulting in the model's inability to generalize.

Figure 8 illustrates the extent to which the predictions are being influenced by the confounders. In Figure 8a, it can be observed that the majority of brains with large volumes are classified as male, while the majority of low-volume brains are categorized as female. A similar trend is seen in Figure 8b, where males are mostly classified as high alcohol users, while females are not. In Figure 8c, the TM test demonstrates how the model predicts around the mean value of the test, with regression values for younger samples mostly falling between 300 and 400 scores.

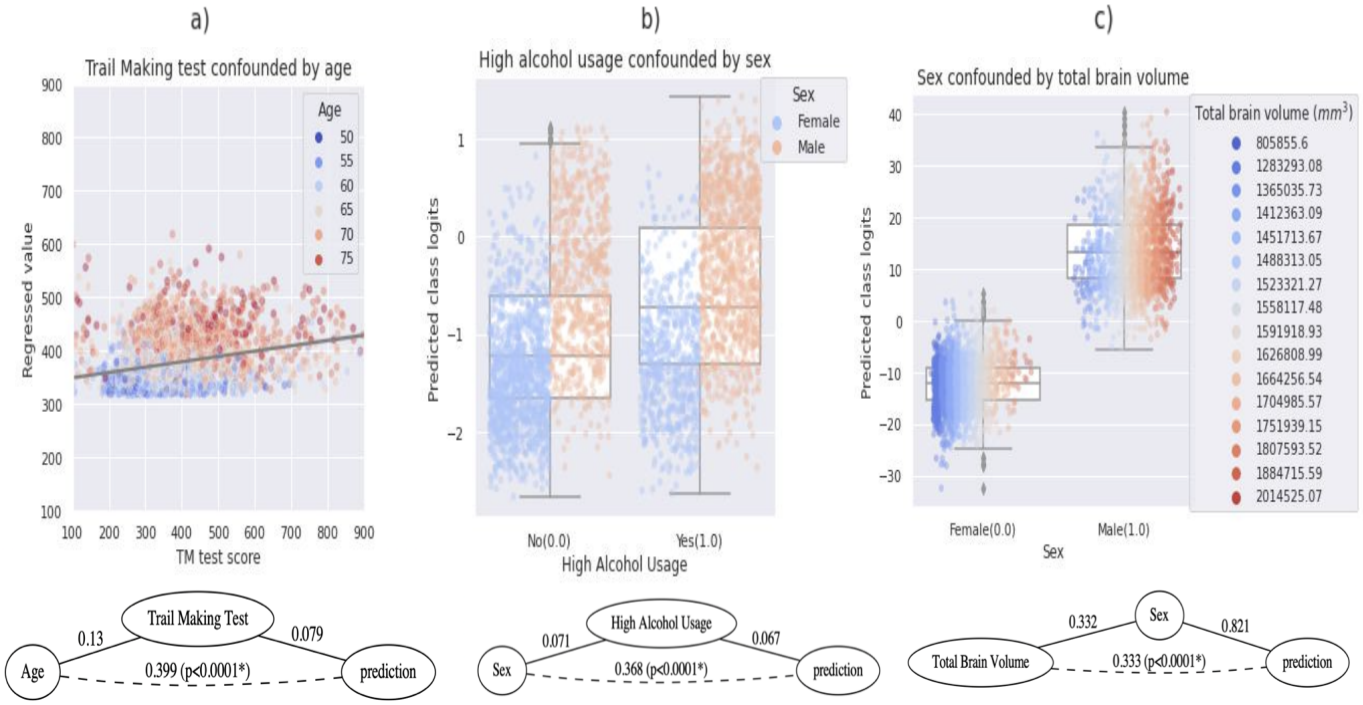

Figure 8: Predictions of the confounded raw models. Image a) presents the regressed values of the Trail Making Test score for subjects in the test set. Images b) and c) exhibit the predicted logits produced by the model for high alcohol usage and sex, respectively. Logits map probability values from 0 to 1 to $-\infty$ to ∞. The color-coding of the confounder variables (a: age, b: sex, c: total brain volume) reveals confounding bias for all phenotypes. Below each image, the corresponding confounded graph is depicted with the R^2 of each variable. This bias is strongly detected by the Partial Confounder Test ($p < 0.0001$).

The issue of confounding effects is not trivial. If an automatic DL algorithm is introduced in a clinical setting to detect potential alcohol misuse disorder and the effect of sex as a confound is not controlled for, the model may discriminate by classifying more males than females as having unhealthy behavior. Furthermore, in order to identify the diagnostic biomarkers of dementia that contribute to poor performance on the TM test and are not due to aging, it is essential to remove the age signal that the model is capturing. For these reasons, addressing these confounds is of crucial significance in order to obtain results that are robust, trustworthy, fair, and generalizable.

3.4 Model's performance after debiasing

After retraining the models with debiasing techniques, the results in terms of performance are shown in Table 5. In sex, there are no significant changes in accuracy across the methods.

For high alcohol usage, the balanced accuracy improved slightly for reweighing (by approximately 1-2%) but significantly improved for the PMDN Layer (85%) when controlling for sex confounding. However, this improvement was not sustained on the test set, as the reported metric only reached 56.45%, lower than the raw model. In the case of the TM test, reweighing's limitation of having to binarize the values makes its accuracy results not comparable. However, in a binary setting, the model appears to have the ability to generalize (8.43% better than chance). For the PMDN Layer, the performance on the validation set improved, but the model still failed to predict on the test set. These results require further evaluation in conjunction with debiasing outcomes.

Table 5: Model's performance before and after the debiasing techniques

		Raw		Reweighing		PMDN Layer	
Phenotype - Confounder		Validation Set	Test Set	Validation Set	Test Set	Validation Set	Test Set
Sex	Total brain volume	0.9982	0.9884	0.9952	0.9888	0.9952	0.9841
High Alcohol Usage	Sex	0.5997	0.5763	0.6096	0.5931	0.85	0.5645
Trail Making Test*	Age	0.086	0.0002	0.614*	0.5843*	0.0958	0.0002

*As the Reweighing technique is limited to only categorical classes, the Trail Making Test results are representative of binary classification.

3.5 Primary confounders analysis after debiasing

To assess the effectiveness of the debiasing methods in controlling for confounders, we evaluate the $R^2_{\hat{y}c}$ and p-value of the PCT again. The results are illustrated in fig.9.

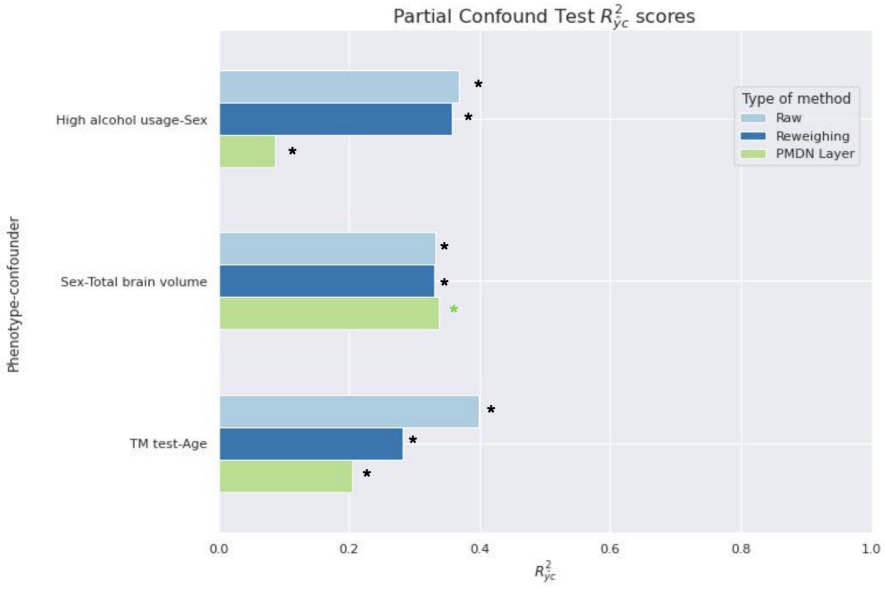

Figure 9: Comparision of the $R^2_{\hat{y}c}$ between methods.The * indicates $p < 0.0001$ and green-colored * $p = 0.01$.

The applied methods were partially effective in eliminating the confounding signal. Despite this, as evidenced by the p-values none of them succeeded in fully controlling for it. Reweighing was ineffective, only reducing confounding by a minor 1-2%. The reweighing values on the TM test are not significant as it was trained on the binary variable, but it can be seen that for binary classification, the bias towards age is weaker. On the other hand, the PMDN Layer mitigated the confounding bias for both sex and age in their respective phenotypes. In control for brain volume, the method did not reduce the $R^2_{\hat{y}c}$, but it did increase the p-value to 1%, indicating significance in controlling for it.

3.5.1 Sex and total brain volume

With regard to the results obtained from the PCT for each method, the effort to reduce bias in the use of total brain volume for sex prediction was not successful. To provide further insight, the influence of debiasing methods on the predicted class logits for the test set is depicted in fig.10. The samples are distinguished by their total brain volume, indicated through the use of color. The graphs below each plot show the coefficients of determination between y, $\hat{y}$, and c.

For the raw model, a bias towards total brain volume is evident, as blue dots (corresponding to lower brain volumes) are predominantly classified as female, while red dots (representing larger brain volumes) are classified as male. Despite the use of reweighing, this pattern remained unchanged, with blue and red dots still accumulating in their respective correlated classes. The reduction in $R^2_{\hat{y}c}$ was only slight, at 0.3%, indicating that the method did not effectively control for any confounding effect. However, the method did produce some effect on the model's output, as the range of values in the predicted logits was reduced.

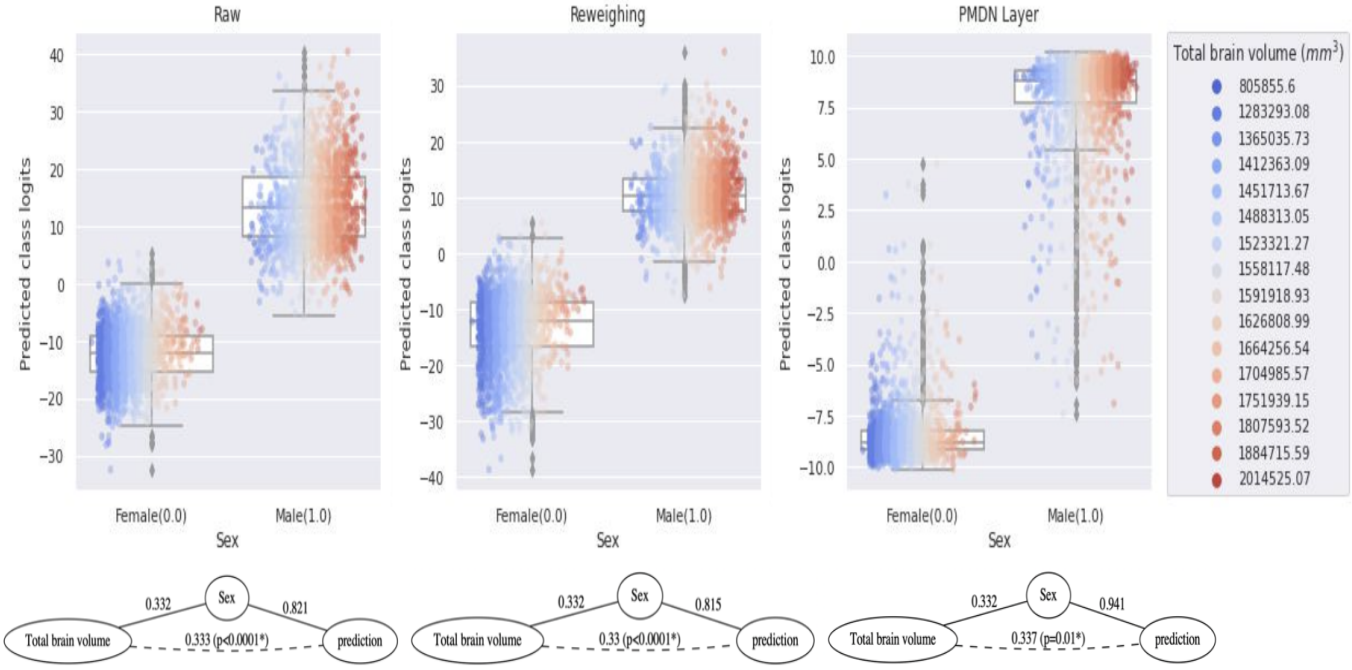

Figure 10: Effect of the debiasing methods controlling for total brain volume in sex prediction. The plots display the predicted logits for each approach. Logits map probability values from 0 to 1 to $-\infty$ to ∞. The color-coding of total brain volume highlights the confounding bias for sex prediction, with male samples being associated with larger brain volume (red) and female samples with smaller brain volume (blue). Although both reweighing and PMDN layer fail to control for the confounder, the PMDN layer reduces the statistical significance of the R^2, demonstrating incipient effective control.

The implementation of the PMDN Layer resulted in a slight increase of 0.4% in $R^2_{\hat{y}c}$, conversely, the p-value increased to 1%. Although this increase is not significant, it does indicate that the method is having a slight debiasing effect. The predicted logits were constrained within a lower and upper bound of -10 to 10, which may be the result of the PMDN Layer's unsuccessful attempt to control for the confound in the model's output. During training, the β learning rate was set to 0.02. This hyperparameter appears to have a direct correlation with the strength with which the PMDN Layer controls confounding factors [21]. Additionally, incorporating additional PMDN Layers into the network architecture would likely further enhance control. However, this falls outside the scope of the current study. As such, further research could adapt these two approaches to determine whether the method is able to control for the confounder.

3.5.2 High alcohol usage and sex

The same analysis has been conducted for high alcohol usage. In fig11, it can be seen Raw and Reweighing have the same distribution of males and females, and almost the same predicted logits, pinpointing the reweighing method failed. On the other hand, the PMDN Layer effectively reduced the bias. Although the bias was not fully eliminated, as evidenced by the p-value being less than 0.01, the distributions of males and females in each class became almost identical. This resulted in a small decrease (1.2%) in balanced accuracy on the test set, as shown in Table 5. This can occur when controlling for confounders, as the model's predictions may be solely based on bias. When the correlation is removed, the model may fail to perform better than chance. Nevertheless, in this case, the method still managed to reduce nearly 30% of the c signal (from 0.3688 to 0.087) and produced significant predictions.

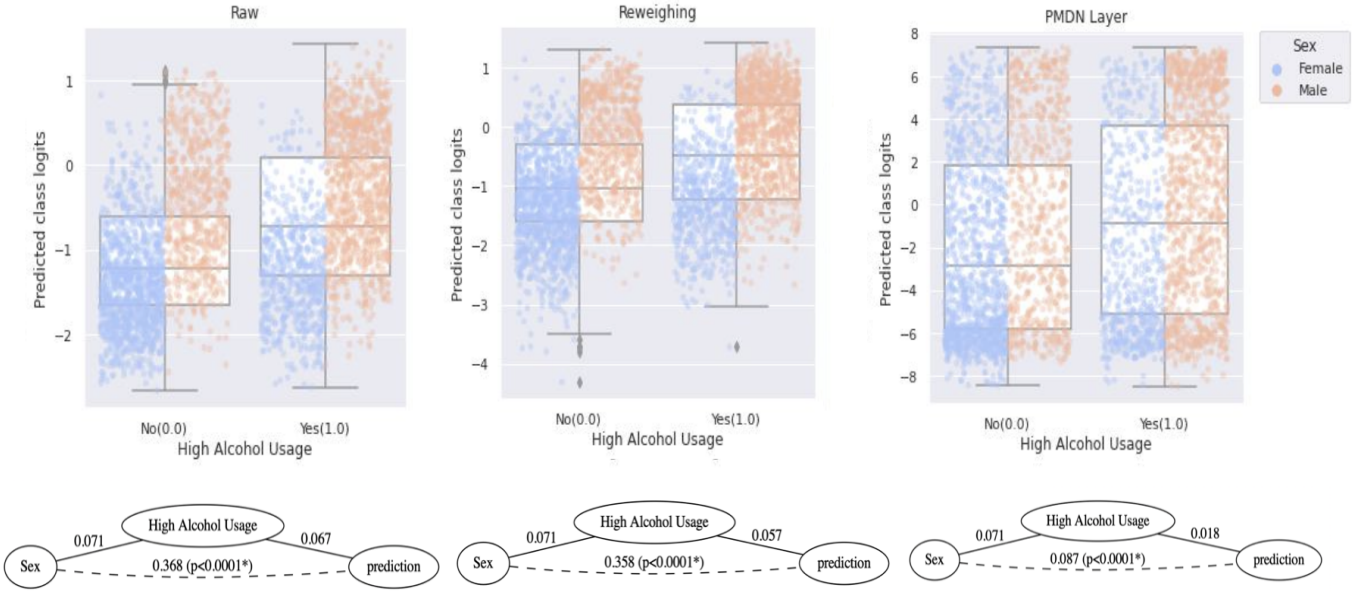

Figure 11: Effect of the debiasing techniques controlling for sex in high alcohol usage prediction. The plots illustrate the predicted class logits for each method. Logits map probability values from 0 to 1 to $-\infty$ to ∞. The color-coding of sex reveals confounding bias for high alcohol usage prediction, high alcohol users are associated with males(red) while non-high alcohol users with females(blue). The PMDN layer significantly controls for the bias, resulting in a more balanced distribution of confound labels.

3.5.3 Trail Making test and age

After following the same procedure for the Trail Making test, the regression values are displayed in fig.12. In the baseline model, TM scores are predicted around the mean value and

with lower scores for younger subjects and higher for older ones. In the reweighing model, although the data is binarized, the bias is still evident, and the $R^2_{\hat{y}c}$ is lower. In the PMDN Layer, the range of predicted scores is wider, indicating an improvement, and the values for age are more equally distributed. This suggests that the method effectively mitigated the influence of age. However, the prediction of the TM test was not successful in the test set, so we cannot conclude that this phenotype can be predicted from T1-weighted MRI scans.

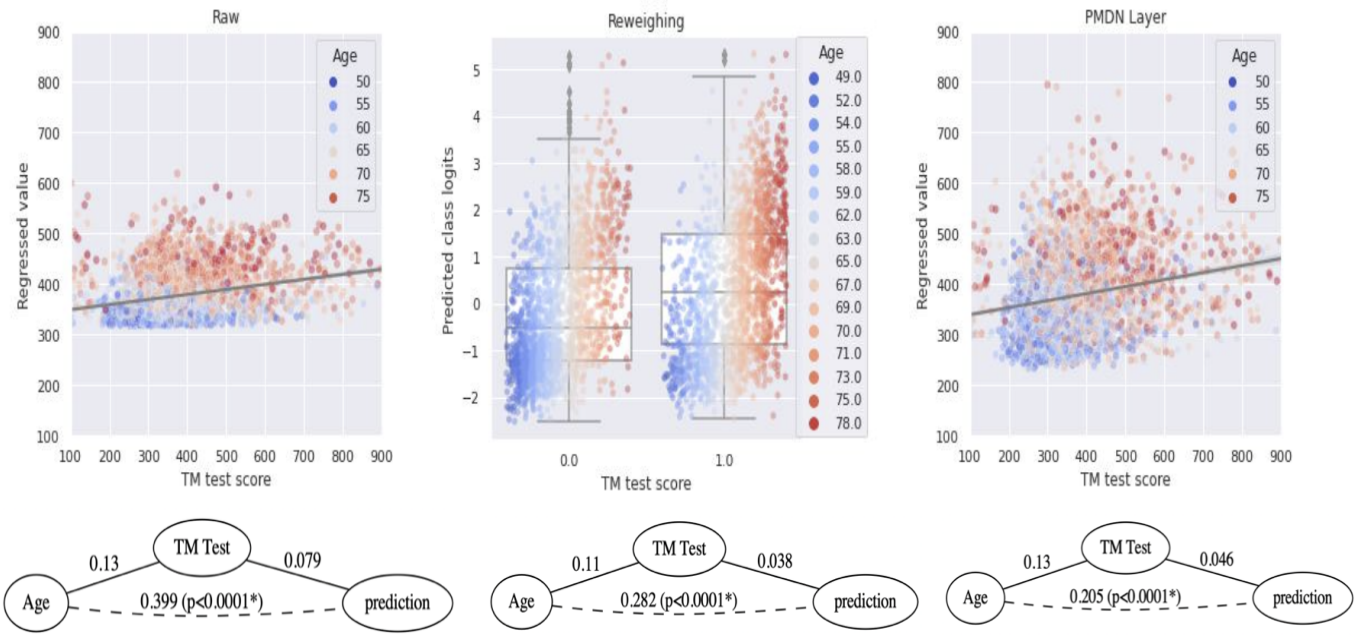

Figure 12: Impact of the Debiasing Methods on Controlling for Age in Trail Making Test Prediction. The plots present the regressed values for each method. The color-coding of age highlights the confounding bias for TM Test prediction, with high scores being associated with older subjects (red) and low scores with younger subjects (blue). In the raw model, score values are predicted around the mean values. Binary classification with reweighing shows a reduced $R^2_{\hat{y}c}$, but still fails to eliminate the age effect. The PMDN layer improves the unbiased predictions with regard to age, although it does not fully control for the confounder ($p < 0.0001$).

Overall, the reweighing approach proved to be ineffective in addressing the confounders that obstruct the prediction of brain phenotypes. On the other hand, the PMDN Layer showed promising results in controlling the model's output. Specifically, it was particularly successful in controlling for age and sex in their respective tasks. Further fine-tuning of the β learning rate for each task or the addition of additional penalty layers may result in the full removal of the confounder's signal.

CHAPTER 4.CONCLUSIONS

4.1 Conclusions

This study presents a framework for systematically investigating the confounding factors that hinder the performance of models in predicting brain phenotypes. The following conclusions have been drawn:

1. The age phenotype can be predicted without controlling for confounding variables.
2. The primary confounds for the prediction of sex and high alcohol usage are brain volume and sex, respectively. As regards their results in PCT and FCT, both have been shown to partially but not completely influence the predictions.
3. The primary confounder for the Trail Making test was age. The model was unable to generalize this brain phenotype prediction in the test set and was fully confounded with age.
4. PCT and FCT set a good combination for quantifying confounding factors in DL models.
5. Reweighing is not a suitable method for controlling confounding factors due to its limitations in classification tasks and poor performance in addressing bias.
6. The PMDN layer has been shown to be an effective method for controlling confounds. Although it does not completely remove the confounds' signal, it has been demonstrated to significantly reduce their influence.

4.2 Future Research

The study of confounders in ML and DL is an emerging field, as evidenced by the novelty of the literature and methods in this area. It is suggested that additional reproducible studies should be conducted to identify and quantify confounding factors when larger neuroimaging datasets are available. There is a need to develop a standardized framework and methods for controlling these confounding factors in DL. The PMDN Layer has produced promising results in this regard, and future experiments should focus on fine-tuning the β learning rate parameter to completely remove the signal of confounding factors.

CHAPTER 5. BIBLIOGRAPHY

[1] A. Fourcade and R. Khonsari, "Deep learning in medical image analysis: A third eye for doctors," *Journal of stomatology, oral and maxillofacial surgery*, vol. 120, no. 4, pp. 279–288, 2019.

[2] T. Wolfers, J. K. Buitelaar, C. F. Beckmann, B. Franke, and A. F. Marquand, "From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics," *Neuroscience & Biobehavioral Reviews*, vol. 57, pp. 328–349, 2015.

[3] A. N. Nielsen, D. M. Barch, S. E. Petersen, B. L. Schlaggar, and D. J. Greene, "Machine learning with neuroimaging: Evaluating its applications in psychiatry," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 5, no. 8, pp. 791–798, 2020.

[4] J. V. Haxby, "Multivariate pattern analysis of fmri: The early beginnings," *Neuroimage*, vol. 62, no. 2, pp. 852–855, 2012.

[5] L. Snoek, S. Miletić, and H. S. Scholte, "How to control for confounds in decoding analyses of neuroimaging data," *Neuroimage*, vol. 184, pp. 741–760, 2019.

[6] F. Eitel *et al.*, "Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional mri using layer-wise relevance propagation," *NeuroImage: Clinical*, vol. 24, p. 102 003, 2019.

[7] R. P. Rane, A. Heinz, and K. Ritter, "Aim in alcohol and drug dependence," *Artificial Intelligence in Medicine*, pp. 1619–1628, 2022.

[8] S. Vieira, X. Liang, R. Guiomar, and A. Mechelli, "Can we predict who will benefit from cognitive-behavioural therapy? a systematic review and meta-analysis of machine learning studies," *Clinical Psychology Review*, p. 102 193, 2022.

[9] A. Dadu *et al.*, "Identification and prediction of parkinson's disease subtypes and progression using machine learning in two cohorts," *npj Parkinson's Disease*, vol. 8, no. 1, pp. 1–12, 2022.

[10] T. He *et al.*, "Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics," *NeuroImage*, vol. 206, p. 116 276, 2020.

[11] G. Starke, E. De Clercq, S. Borgwardt, and B. S. Elger, "Computing schizophrenia: Ethical challenges for machine learning in psychiatry," *Psychological Medicine*, vol. 51, no. 15, pp. 2515–2521, 2021.

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

[13] F. Eitel, M.-A. Schulz, M. Seiler, H. Walter, and K. Ritter, "Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research," *Experimental Neurology*, vol. 339, p. 113 608, 2021.

[14] R. P. Rane *et al.*, "Structural differences in adolescent brains can predict alcohol misuse," *Elife*, vol. 11, e77545, 2022.

[15] Z. Xu *et al.*, "A survey of fairness in medical image analysis: Concepts, algorithms, evaluations, and challenges," *arXiv preprint arXiv:2209.13177*, 2022.

[16] D. Chyzhyk, G. Varoquaux, B. Thirion, and M. Milham, "Controlling a confound in predictive models with a test set minimizing its effect," in *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, IEEE, 2018, pp. 1–4.

[17] E. Thibeau-Sutre, B. Couvy-Duchesne, D. Dormont, O. Colliot, and N. Burgos, "Mri field strength predicts alzheimer's disease: A case example of bias in the adni data set," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2022, pp. 1–4.

[18] M. Klingenberg *et al.*, "Higher performance for women than men in mri-based alzheimer's disease detection," 2022.

[19] R. Dinga, L. Schmaal, B. W. Penninx, D. J. Veltman, and A. F. Marquand, "Controlling for effects of confounding variables on machine learning predictions," *BioRxiv*, 2020.

[20] E. Puyol-Antón *et al.*, "Fairness in cardiac mr image analysis: An investigation of bias due to data imbalance in deep learning based segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 413–423.

[21] A. Vento, Q. Zhao, R. Paul, K. M. Pohl, and E. Adeli, "A penalty approach for normalizing feature distributions to build confounder-free models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, pp. 387–397.

[22] C. Sudlow *et al.*, "Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, vol. 12, no. 3, e1001779, 2015.

[23] T. J. Littlejohns *et al.*, "The uk biobank imaging enhancement of 100,000 participants: Rationale, data collection, management and future directions," *Nature communications*, vol. 11, no. 1, pp. 1–12, 2020.

[24] F. Alfaro-Almagro *et al.*, "Confound modelling in uk biobank brain imaging," *NeuroImage*, vol. 224, p. 117 002, 2021.

[25] F. Alfaro-Almagro *et al.*, "Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank," *Neuroimage*, vol. 166, pp. 400–424, 2018.

[26] R. M. Reitan, "Validity of the trail making test as an indicator of organic brain damage," *Perceptual and motor skills*, vol. 8, no. 3, pp. 271–276, 1958.

[27] M.-A. Schulz, D. Bzdok, S. Haufe, J.-D. Haynes, and K. Ritter, "Performance reserves in brain-imaging-based phenotype prediction," *bioRxiv*, 2022.

[28] S. More, S. B. Eickhoff, J. Caspers, and K. R. Patil, "Confound removal and normalization in practice: A neuroimaging based sex prediction case study," in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*, Springer, 2021, pp. 3–18.

[29] R. I. Scahill *et al.*, "A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging," *Archives of neurology*, vol. 60, no. 7, pp. 989–994, 2003.

[30] C. R. Bowie and P. D. Harvey, "Administration and interpretation of the trail making test," *Nature protocols*, vol. 1, no. 5, pp. 2277–2281, 2006.

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

[32] D. P. Kingma, "&Ba j.(2014). adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015.

[33] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.

[34] B. Abdollahi, N. Tomita, and S. Hassanpour, "Data augmentation in training deep learning models for medical image analysis," in *Deep learners and deep learner descriptors for medical applications*, Springer, 2020, pp. 167–180.

[35] P. Chlap *et al.*, "A review of medical image data augmentation techniques for deep learning applications," *Journal of Medical Imaging and Radiation Oncology*, vol. 65, no. 5, pp. 545–563, 2021.

[36] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[37] Z. Wang *et al.*, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8919–8928.

[38] N. Joshi and P. Burlina, "Ai fairness via domain adaptation," *arXiv preprint arXiv:2104.01109*, 2021.

[39] Q. Zhao, E. Adeli, and K. M. Pohl, "Training confounder-free deep learning models for medical applications," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[40] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni, "On the fairness of privacy-preserving representations in medical applications," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, Springer, 2020, pp. 140–149.

[41] S. Du, B. Hers, N. Bayasi, G. Hamarneh, and R. Garbi, "Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning," *arXiv preprint arXiv:2208.10013*, 2022.

[42] Y. Zhou *et al.*, "Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr," *arXiv preprint arXiv:2111.11665*, 2021.

[43] Y. Wu, D. Zeng, X. Xu, Y. Shi, and J. Hu, "Fairprune: Achieving fairness through pruning for dermatological disease diagnosis," *arXiv preprint arXiv:2203.02110*, 2022.

[44] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[45] R. K. Bellamy *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[46] T. Beuzen, L. Marshall, and K. D. Splinter, "A comparison of methods for discretizing continuous variables in bayesian networks," *Environmental modelling & software*, vol. 108, pp. 61–66, 2018.

[47] M. Lu *et al.*, "Metadata normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 917–10 927.

[48] T. Spisak, "Statistical quantification of confounding bias in machine learning models," *GigaScience*, vol. 11, 2022.

A. ANNEX. GitHub Repository

The code that has been created constitutes the backbone of this project, as it outlines a significant portion of the procedures that comprise the methodology developed to reach its goals. The code has been separated into two directories, each of which encompasses the pipelines for raw data training and reweighing, as well as the PMDN layer approach. These two directories can be located in the GitHub repository indicated below:

`https://github.com/GonzaloCardenalAl/thesis`

- Directory **raw_and_reweighing**: contains the pipeline to train the raw ResNet50 and the reweighing approach. Additionally, it contains the notebooks and related python files to process and create the training and test set files.
- Directory **PMDN**: contains the pipeline to train the PMDNResNet50 model.
- Directory **nitorch**: contains files needed to run the 'CNNpipeline.py' such as the trainer or the inference function.

B. ANNEX. Complementary Figures

B.1 Data distribution for the different variables

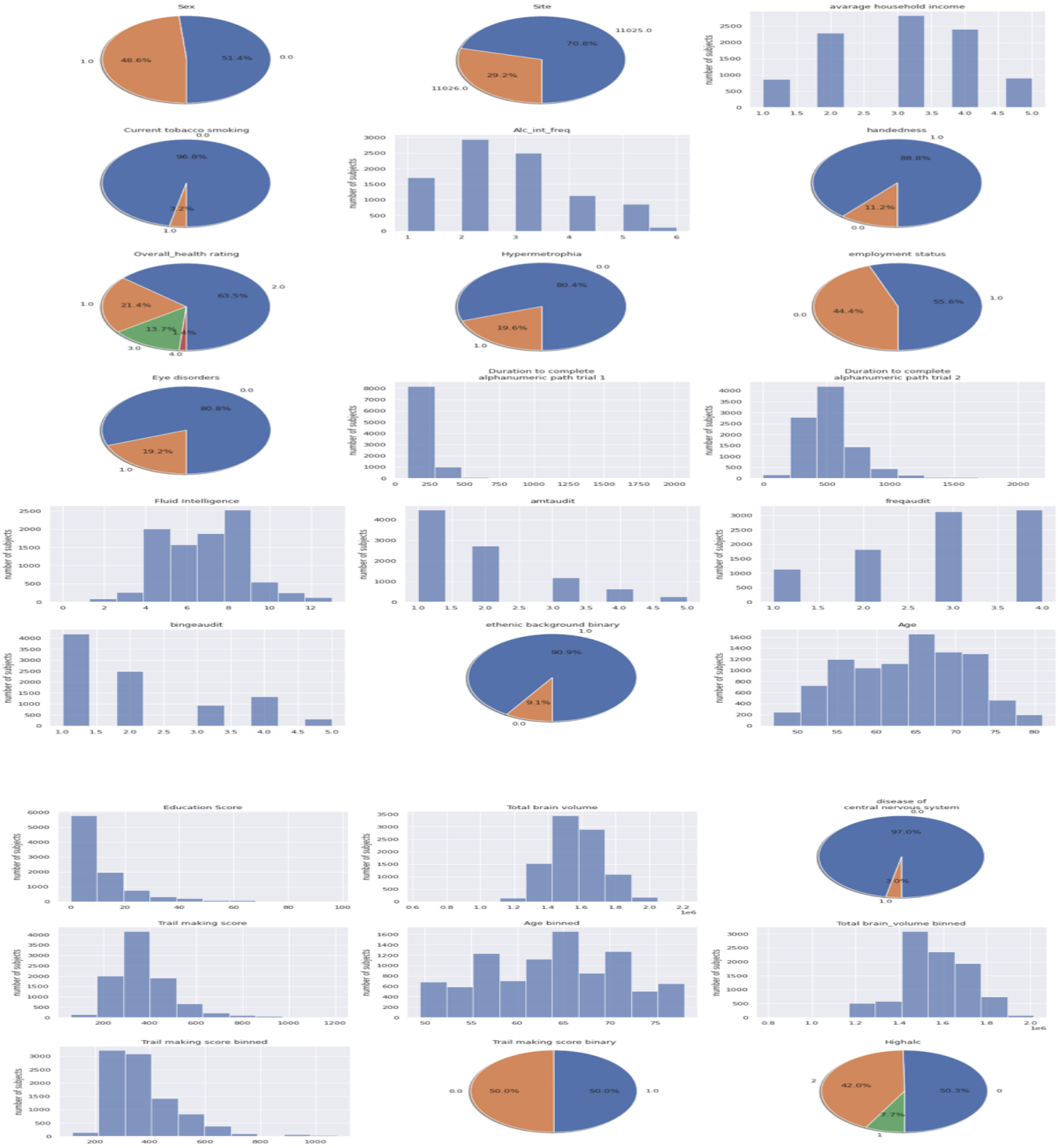

Figure 13: Distribution of the training set for each variable.

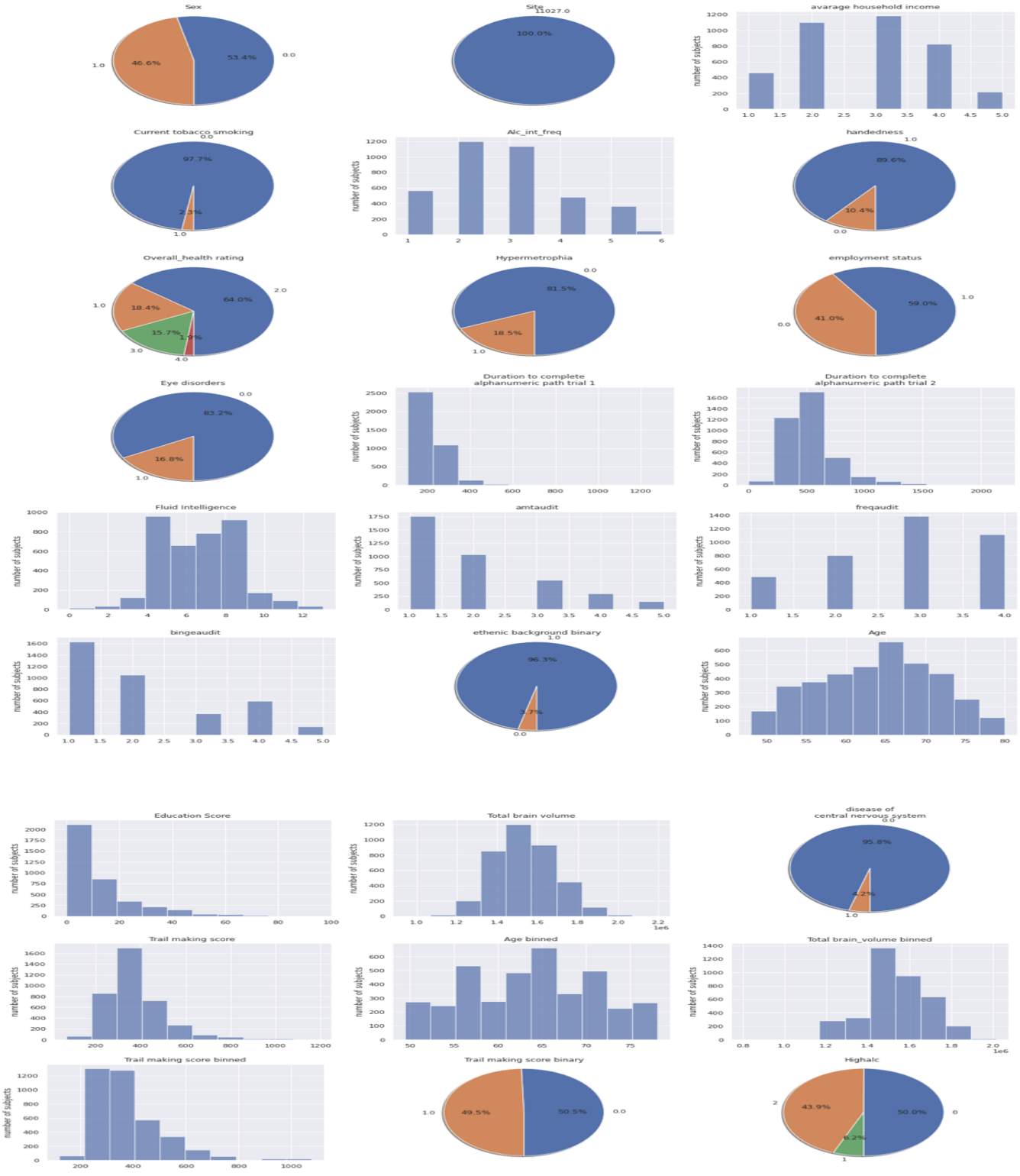

Figure 14: Distribution of the hold set for each variable.

B.2 Training Curves

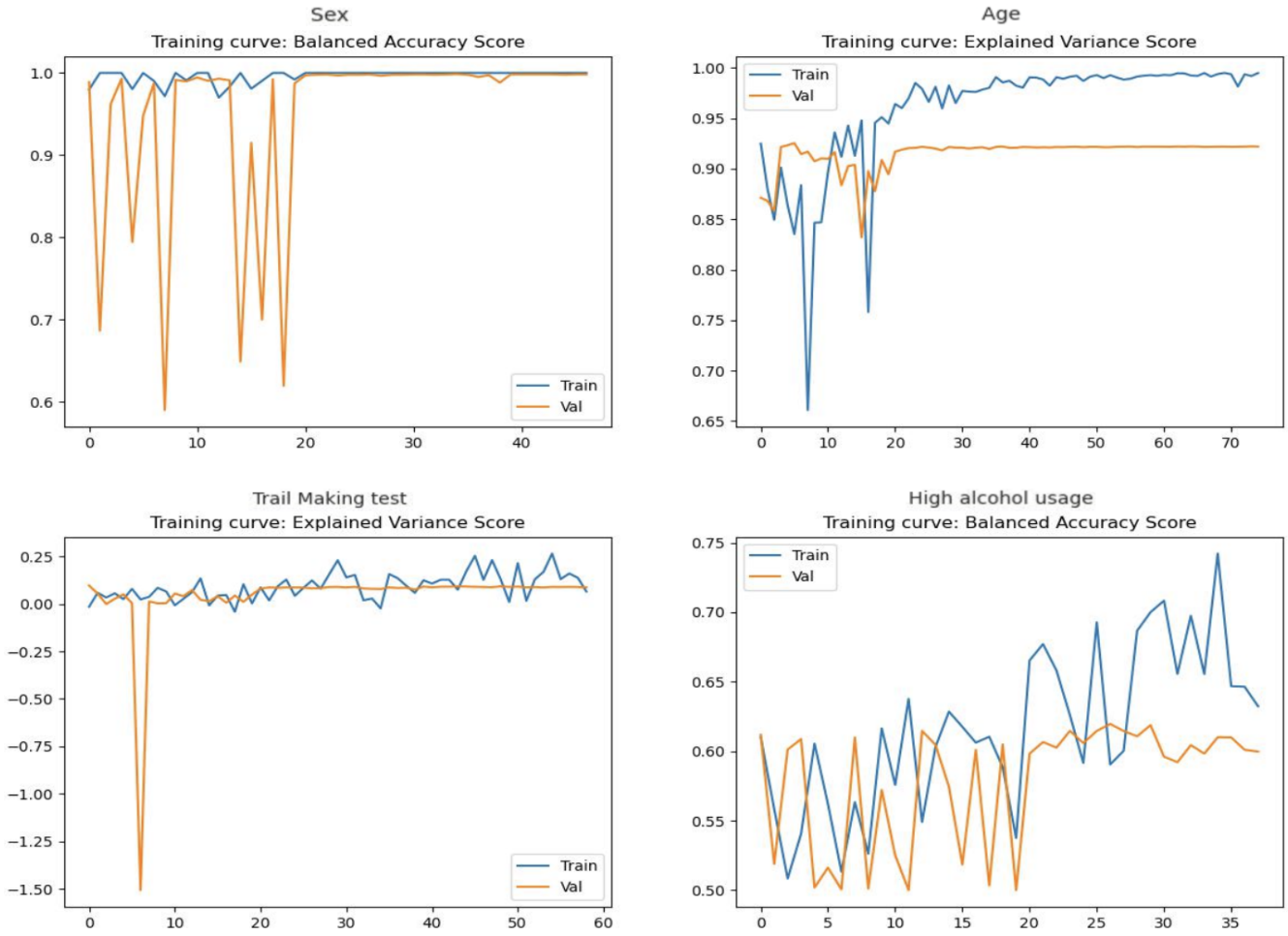

Figure 15: Training curves of the raw model for each phenotype.

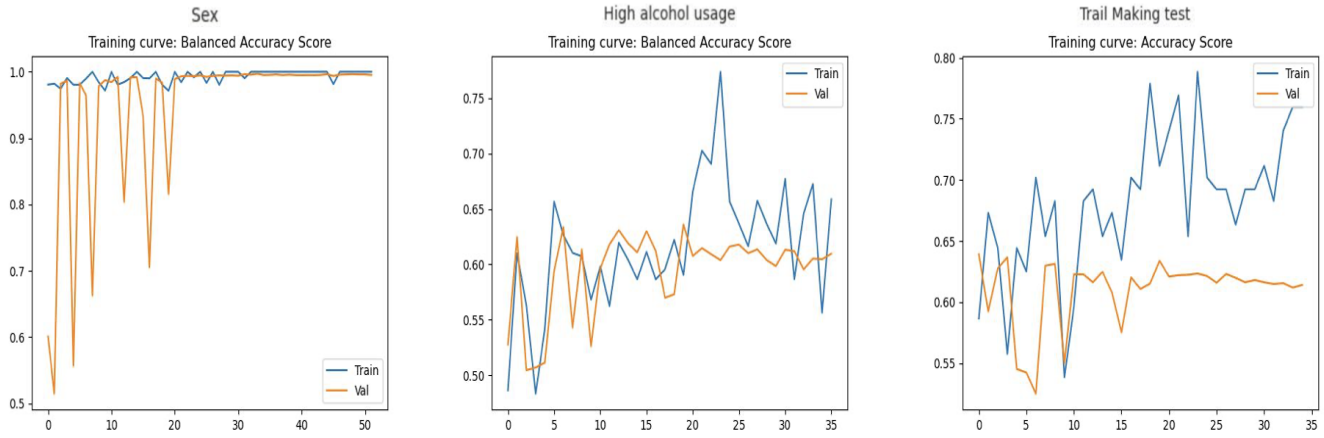

Figure 16: Training curves of reweighing for each confounded phenotype.

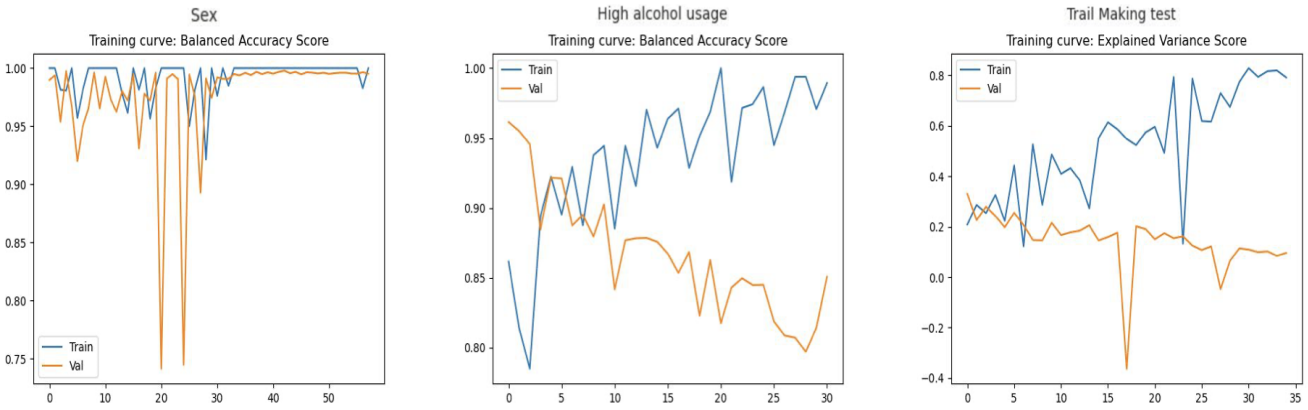

Figure 17: Training curves of PMDN for each confounded phenotype.