Statistical interpretation of high-dimensional complex prediction models for biomedical data

Ahmad Chamma

université PARIS-SACLAY

Statistical Interpretation of High-Dimensional Complex Prediction Models for Biomedical Data

Interprétation statistique des modèles de prédiction complexes à haute dimension pour les données biomédicales

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580 : Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat: Sciences du traitement du signal et des images
Graduate School : Informatique et sciences du numérique. Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Inria Saclay-Île-de-France (Université Paris-Saclay, Inria)**, sous la direction de **Bertrand THIRION**, Directeur de recherche, et sous le co-encadrement de **Denis ENGEMANN**, Biomarker & Experimental Medicine Leader à Roche pRED, Basel, Suisse

Thèse soutenue à Paris-Saclay, le 14 Juin 2024, par

Ahmad CHAMMA

Composition du jury

Membres du jury avec voix délibérative

Erwan SCORNET Professeur des universités, Sorbonne Université, France	Président
Anne-Laure BOULSETEIX Professeur des universités, Ludwig Maximilian University of Munich, Allemagne	Rapporteur & Examinatrice
Moritz GROSSE-WENTRUP Professeur des universités, University of Vienna, Autriche	Rapporteur & Examinateur
Sylvain CHEVALIER Professeur des universités, Université Paris-Saclay, France	Examinateur

THESE DE DOCTORAT

NNT : 2024UPASG028

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title: Statistical Interpretation of High-Dimensional Complex Prediction Models for Biomedical Data

Keywords: Interpretability, Machine learning, Statistics, Deep learning

Abstract: Modern large health datasets represent population characteristics in multiple modalities, including brain imaging and socio-demographic data. These large cohorts make it possible to predict and understand individual outcomes, leading to promising results in the epidemiological context of forecasting/predicting the occurrence of diseases, health outcomes, or other events of interest. As data collection expands into different scientific domains, such as brain imaging and genomic analysis, variables are related by complex, possibly non-linear dependencies, along with high degrees of correlation. As a result, popular models such as linear and tree-based techniques are no longer effective in such high-dimensional settings. Powerful non-linear machine learning algorithms, such as Random Forests (RFs) and Deep Neural Networks (DNNs), have become important tools for characterizing inter-individual differences and predicting biomedical outcomes, such as brain age. Explaining the decision process of machine learning algorithms is crucial both to improve the performance of a model and to aid human understanding. This can be achieved by assessing the importance of variables. Traditionally, scientists have favored simple, transparent models such as linear regression, where the importance of variables can be easily measured by coefficients. However, with the use of more advanced methods, direct access to the internal structure has become limited and/or uninterpretable from a human perspective. As a result, these methods are often referred to as "black box" methods. Standard approaches based on Permutation Importance (PI) assess the importance of a variable by measuring the decrease in the loss score when the variable of interest is replaced by its permuted version. While these approaches increase the transparency of black-box models and provide statistical validity, they can produce unreliable importance assessments when variables are correlated.

The goal of this work is to overcome the limitations of standard permutation importance by integrating conditional schemes. Therefore, we investigate two model-agnostic frameworks, Conditional Permutation Importance (CPI) and Block-Based Conditional Permutation Importance (BCPI), which effectively account for correlations between covariates and overcome the limitations of PI. We present two new algorithms designed to handle situations with correlated variables, whether grouped or ungrouped. Our theoretical and empirical results show that CPI provides computationally efficient and theoretically sound methods for evaluating individual variables. The CPI framework guarantees type-I error control and produces a concise selection of significant variables in large datasets.

BCPI presents a strategy for managing both individual and grouped variables. It integrates statistical clustering and uses prior knowledge of grouping to adapt the DNN architecture using stacking techniques. This framework is robust and maintains type-I error control even in scenarios with highly correlated groups of variables. It performs well on various benchmarks. Empirical evaluations of our methods on several biomedical datasets showed good face validity. Our methods have also been applied to multimodal brain data in addition to socio-demographics, paving the way for new discoveries and advances in the targeted areas. The CPI and BCPI frameworks are proposed as replacements for conventional permutation-based methods. They provide improved interpretability and reliability in estimating variable importance for high-performance machine learning models.

Titre: Interprétation statistique des modèles de prédiction complexes à haute dimension pour les données biomédicales

Mots clés: Interpretabilité, Apprentissage, Statistiques, Apprentissage approfondi

Résumé: Les grands jeux de données de santé produits, qui représentent les caractéristiques de la population selon de multiples modalités, permettent de prédire et de comprendre les résultats individuels. À mesure que la collecte de données s'étend aux domaines scientifiques, tels que l'imagerie cérébrale, les variables sont liées par des dépendances complexes, éventuellement non linéaires, ainsi que par des degrés élevés de corrélation. Par conséquent, les modèles populaires tels que les techniques linéaires et à base d'arbres de décision ne sont plus efficaces dans ces contextes à haute dimension. De puissants algorithmes d'apprentissage automatique non linéaires, tels que les forêts aléatoires et les réseaux de neurones profonds, sont devenus des outils importants pour caractériser les différences interindividuelles et prédire les résultats biomédicaux, tels que l'âge du cerveau. Il est essentiel d'expliquer le processus de décision des algorithmes d'apprentissage automatique, à la fois pour améliorer les performances d'un modèle et pour faciliter la compréhension. Cet objectif peut être atteint en évaluant l'importance des variables. Traditionnellement, les scientifiques ont privilégié des modèles simples et transparents tels que la régression linéaire, où l'importance des variables peut être facilement mesurée par des coefficients. Cependant, avec l'utilisation de méthodes plus avancées, l'accès direct à la structure interne est devenu limité et/ou ininterprétable d'un point de vue humain. C'est pourquoi ces méthodes sont souvent appelées méthodes "boîte noire". Les approches standard basées sur l'importance par permutation (PI) évaluent l'importance d'une variable en mesurant la diminution du score de perte lorsque la variable d'intérêt est remplacée par sa version permutée. Bien que ces approches augmentent la transparence des modèles de boîte noire et offrent une validité statistique, elles peuvent produire des évaluations d'importance peu fiables lorsque les variables sont corrélées.

L'objectif de ce travail est de surmonter les limites de l'importance de permutation standard en intégrant des schémas conditionnels. Par conséquent, nous développons deux cadres génériques, l'importance par permutation conditionnelle (CPI) et l'importance par permutation conditionnelle basée sur des blocs (BCPI), qui prennent efficacement en compte les corrélations entre les covariables et surmontent les limites de l'importance par permutation. Nous présentons deux nouveaux algorithmes conçus pour traiter les situations où les variables sont corrélées, qu'elles soient groupées ou non. Nos résultats théoriques et empiriques montrent que CPI fournit des méthodes efficaces sur le plan du calcul et solides sur le plan théorique pour l'évaluation des variables individuelles. Le cadre de CPI garantit le contrôle des erreurs de type-I et produit une sélection concise des variables significatives dans les grands ensembles de données.

BCPI présente une stratégie de gestion des variables individuelles et groupées. Elle intègre le regroupement statistique et utilise la connaissance préalable du regroupement pour adapter l'architecture du réseau DNN à l'aide de techniques d'empilement. Ce cadre est robuste et maintient le contrôle de l'erreur de type-I même dans des scénarios avec des groupes de variables fortement corrélées. Il donne de bons résultats sur divers points de référence. Les évaluations empiriques de nos méthodes sur plusieurs jeux de données biomédicales ont montré une bonne validité apparente. Nous avons également appliqué ces méthodes à des données cérébrales multimodales ainsi qu'à des données sociodémographiques, ouvrant la voie à de nouvelles découvertes et avancées dans les domaines ciblés. Les cadres CPI et BCPI sont proposés en remplacement des méthodes conventionnelles basées sur la permutation. Ils améliorent l'interprétabilité de l'estimation de l'importance des variables pour les modèles d'apprentissage à haute performance.

Acknowledgments

First, I am profoundly grateful to **God** for His unwavering guidance and grace, which have enabled me to successfully complete my thesis.

Secondly, I am very grateful to my supervisors, **Bertrand Thirion** and **Denis Engemann** for always being available at any time and for their valuable guidance. I can't thank them enough for their support during the difficult times encountered in stressful periods.

Furthermore, I would like to thank the two reviewers, Prof. **Anne-Laure Boulseteix** and Prof. **Moritz Grosse-Wentrup** who kindly accepted to read my dissertation and to deliver a report evaluating my work.

Then, I want to thank all the great jury members who kindly accepted to be in my Ph.D. defense committee, Prof. **Sylvain Chevalier** and Prof. **Erwan Cornet**.

As for my colleagues on the great Inria team, words can't express how grateful I am for the beautiful memories and the fruitful discussions we've shared together. I will never forget the super ping pong games with my colleague **Alexandre Blain** during the breaks.

I will not forget to thank my friends: **Hadi Abdine**, **Mohamad Al Assaad**, **Mohamad Al Sayegh**, **Moussa Kamaleddine** and **Omar Ahmad**, who were next to me during the whole period of my Ph.D. These guys are examples of what we call "true friends", helping, supporting and empowering.

Finally, my most enormous thanks go to my beloved family: **my father**, **mother**, **brothers**, **sisters**, **niece** and **nephews**. These people are the most beautiful blessing I have had, and I will ever have. I would fairly say that without their moral support, I could have surrendered without completing this dissertation. I found them by my side at every turn, encouraging and helping me rise again. I dedicate this thesis to **my beloved** mother, whose unwavering support, encouragement, and belief in me have been my greatest source of inspiration throughout this journey. This dream has blossomed thanks to your nurturing guidance and support.

As I conclude this final section of my thesis chapter, I find myself unable to fully express how this journey has transformed me into a better version of myself.

Ahmad CHAMMA
Palaiseau, June 2024

Contents

List of Figures viii

List of Tables ix

Acronyms xi

Preliminaries 1

Synthèse en français 3
1 Les Schémas Conditionnels, le Remède aux *Fausses* Variables Significatives 4
2 Le Regroupement est la Clé des Paramètres à *Haute-Dimension* 5
3 Importance des Variables pour les Applications de Neuro-imagerie 6
4 Conclusion 6
5 Autres Travaux 6
6 Logiciel 7

Overview 9
1 Conditional Schemes, The Remedy for *Fake* Significant Variables 10
2 Grouping is The Key for *High-Dimensional* Settings 11
3 Variable Importance for Neuroimaging Applications 11
4 Conclusion 12
5 Other Works 12
6 Software 12

I Background 13

1 Motivation and Problem Statement 15
1.1 Machine Learning & High-Dimensional Settings 15
1.2 Single Variable Importance 17
1.3 Hypothesis Testing 19
1.4 Statistical Control 20
1.5 Conditional Inference 21
1.6 Group Variable Importance & Stacking 23
1.7 Problem Setting 24
1.7.1 Single Case 24
1.7.2 Group Case 24
1.8 Conclusion 24

2 Variable Importance Methods **25**
2.1 Non-perturbative Sensitivity Analysis . . . 26
2.1.1 Model Specific Methods . . . 26
2.1.2 Model Agnostic Methods . . . 28
2.2 Removal-based Methods . . . 29
2.2.1 Model Specific Methods . . . 29
2.2.2 Model Agnostic Methods - Instance based . . . 32
2.2.3 Model Agnostic Methods - Population level . . . 34
2.3 Conclusion . . . 36

3 Group Variable Importance Methods **39**
3.1 Variables Grouping . . . 40
3.1.1 Data-driven grouping (Clustering Methods) . . . 40
3.1.2 Knowledge-driven grouping . . . 41
3.2 Solo Group Representative . . . 41
3.2.1 Cluster Summarization via Aggregation . . . 41
3.2.2 Stacking Approach . . . 41
3.3 Grouped Non-perturbative Methods . . . 42
3.4 Grouped Removal-based Methods . . . 43
3.4.1 Grouped Shapley Values . . . 43
3.4.2 Grouped Permutation-based Methods . . . 44
3.4.3 Grouped Refitting-based Methods . . . 45
3.5 Conclusion . . . 45

4 Variable Importance for Population Imaging in Neuroimaging **47**
4.1 Brain Imaging Modalities (Neuroimaging) . . . 47
4.1.1 Magnetic Resonance Imaging (MRI) . . . 48
4.1.2 Functional Magnetic Resonance Imaging (fMRI) . . . 49
4.1.3 Electrophysiological Methods (M/EEG) . . . 49
4.1.4 Combining brain imaging modalities . . . 50
4.2 Extending the reach of neuroscientific research with interpretable machine learning . . . 50
4.3 Proxy Measures . . . 53
4.4 Conclusion . . . 54
4.5 Scientific goals of the thesis . . . 54

II Contributions **55**

Preliminaries **57**

5 Statistical Valid Importance: the Case of Single Variables **59**
5.1 Permutation importance and its limitations . . . 59
5.1.1 The *permutation* approach leads to false detections in the presence of correlations 59
5.2 *Conditional sampling*-based feature importance . . . 61

5.2.1 Main result . . . 61
5.2.2 Conditional Permutation Importance (CPI) Wald statistic asymptotically controls type-I errors: hypotheses, theorem and proof . . . 62
5.2.3 Practical estimation . . . 64
5.3 Experiments & Results . . . 66
5.3.1 Experiment 1: Type-I error control and accuracy when increasing variable correlation . . . 66
5.3.2 Experiment 2: Performance across different settings . . . 66
5.3.3 Experiment 3: Performance benchmark across methods . . . 69
5.3.4 Experiment 4: *Permfit-DNN* vs *CPI-DNN* on Real Dataset UKBB . . . 72
5.4 Discussion . . . 77
5.5 Additional Experiments . . . 78
5.5.1 Exp. 2 - Computational scaling of *CPI-DNN* and leanness . . . 78
5.5.2 Exp. 4 - Large scale simulations . . . 79
5.5.3 Exp. 4 - Age prediction from brain activity (MEG) in Cam-CAN dataset . . . 80
5.5.4 Compare *CPI*'s constructions: *Residuals* vs *Sampling* . . . 82
5.5.5 Practical validation of the normal distribution assumption . . . 82
5.5.6 Random Forest for modeling the conditional distribution and resulting calibration 83

6 Sampling of Continuous, Ordinal and Nominal Correlated Variables 85
6.1 Background and Challenges . . . 85
6.2 Sampling of *Continuous* and *Ordinal* variables . . . 85
6.3 Sampling of *Nominal* variables . . . 86
6.4 Illustrative example . . . 87

7 Statistical Valid Importance: the Case of Grouped Variables 89
7.1 Block-Based Conditional Permutation Importance (BCPI) . . . 89
7.1.1 Define more notations for the groups . . . 89
7.1.2 Group conditional variable importance . . . 89
7.2 *Internal* Stacking Approach . . . 91
7.3 Experiments . . . 92
7.3.1 Experiment 1: Benchmark of grouping methods . . . 92
7.3.2 Experiment 2: Impact of Stacking . . . 94
7.3.3 Experiment 3: Age prediction with UKBB . . . 94
7.4 Results . . . 96
7.5 Discussion . . . 97
7.6 Additional Experiments . . . 100
7.6.1 Exp. 1 - Power & Computation time . . . 100
7.6.2 Exp. 1 - AUC score for *Grouped Shapley* values . . . 101
7.6.3 Exp. 1 - AUC score & Type-I error (Non linear case) . . . 101
7.6.4 Exp. 1 - Power & Computation time (Non linear case) . . . 102
7.6.5 Exp. 2 - Groups with different cardinalities . . . 103
7.6.6 Calibration of p-values between *BCPI-DNN* and *BPI-DNN* . . . 103

7.6.7 Impact of multi-output neurons on variable importance . . . 104
7.7 Pre-defined groups in UK BioBank . . . 105

8 Applications to Brain Imaging **107**
8.1 Exploring the influence of multimodal heterogeneous data on biomedical outcome prediction . . . 107
8.1.1 Challenge: inference of significant multimodal heterogeneous data . . . 107
8.1.2 Study & Results . . . 109
8.2 Multimodal brain data: illuminating age prediction insights . . . 109
8.2.1 Challenge: inference of significant brain imaging modalities . . . 109
8.2.2 Processing pipelines . . . 111
8.2.3 Study & Results . . . 112
8.3 Unlocking age prediction: insights from cortical brain regions . . . 112
8.3.1 Challenge: detection of predictive brain regions . . . 112
8.3.2 Study & Results . . . 114
8.4 Significant frequencies bands for age prediction . . . 114
8.4.1 Challenge: inference of the important frequencies from EEG models . . . 114
8.4.2 Dataset description . . . 116
8.4.3 Processing pipeline . . . 116
8.4.4 Study & Results . . . 116
8.5 Significant frequencies for identifying the status of the eyes . . . 117
8.5.1 Challenge: inferring significant frequency contributions in EEG prediction models 117
8.5.2 Dataset description . . . 117
8.5.3 Processing pipeline . . . 119
8.5.4 Study & Results . . . 119
8.6 Conclusion . . . 120

9 Conclusion **121**

References **145**

List of Figures

1.1 Conditional Inference in Neuroscience . . . 22

4.1 Segmentation of the brain . . . 48
4.2 Segmentation of the brain . . . 51
4.3 Whole-brain functional connectivity data . . . 52

5.1 Construction of CPI . . . 61
5.2 CPI-DNN vs Permfit-DNN . . . 67
5.3 Model comparisons across data-generating scenarios . . . 68
5.4 Extended model comparisons . . . 70
5.5 Extended model comparisons-noPval . . . 71
5.6 Extended model comparisons-Computation time . . . 72
5.7 Evaluation of predictive performance . . . 73
5.8 Real-world empirical benchmark . . . 74
5.3.2-S1 *CPI-DNN* vs *LOCO-DNN* . . . 78
5.3.4-S1 Semi-simulation with UK Biobank . . . 79
5.3.4-S2 Large scale simulation . . . 79
5-SE1 Age prediction from brain activity . . . 80
5-SE2 *Residuals* vs *Sampling* . . . 82
5-SE3 Normal distribution assumption . . . 83
5-SE5 Random forest calibration . . . 83
5-SE4 *CPI-DNN* vs *Permfit-DNN* p-values calibration . . . 84

6.2.1 Category specification of the latent variable with distribution equality . . . 86
6.4.2 Correlation-Adjusted Sampling . . . 88

7.1.1 Block-Based Conditional Permutation Importance . . . 90
7.3.2 Benchmarking grouping methods . . . 92
7.3.3 Impact of Stacking . . . 94
7.3.4 Brain Age prediction in UKBB . . . 95
7.3.1-S1 Exp. 1 - Power & Computation time . . . 100
7.3.1-S2 *Grouped Shapley* values . . . 101
7.3.1-S3 AUC/Type-I Error-Non Linearity . . . 101
7.3.1-S4 Power/Computation Time-Non Linearity . . . 102
7.3.2-S1 Groups of different cardinalities . . . 103
7-SE1 P-values calibration for groups . . . 103
7-SE2 Brain Age prediction in UKBB with multi-output neurones . . . 104

8.1.1 Biomedical outcome prediction in UK Biobank . . . 108
8.2.2 Age prediction in Cam-Can from Brain imaging modalities . . . 110

8.3.3 Brain regions/parcels significantly contributing to age prediction 113
8.4.4 Significant frequencies for age prediction under Riemannian/SPoC projectors in TUAB 115
8.5.5 Significant frequencies for eyes' status under Riemannian/SPoC projectors in LEMON 118

List of Tables

2.1 Summarizing table for single-based methods 26

3.1 Summarizing table for group-based methods 42

7.1 Knowledge-based groups in UK BioBank 105

Acronyms

BCPI Block-Based Conditional Permutation Importance.

BPI Block-Based Permutation Importance.

Cam-CAN Cambridge Center for Ageing and Neuroscience.

CPI Conditional Permutation Importance.

CRT Conditional Randomization Testing.

EEG Electroencephalography.

FDR False Discovery Rate.

fMRI Functional Magnetic Resonance Imaging.

GOPFI Grouped Only Permutation Feature Importance.

GPFI Grouped Permutation Feature Importance.

LEMON Leipzig Study for Mind-Body-Emotion Interactions.

LOGI Leave One Group In.

LOGO Leave One Group Out.

MCMC Markov Chain Monte Carlo.

MDI Mean Decrease Impurity.

ML Machine Learning.

MRI Magnetic Resonance Imaging.

OOB Out-of-bag.

PI Permutation Importance.

RF Random Forest.

ROI Region of Interest.

SPoC Source Power Comodulation.

VI Variable Importance.

Preliminaries

Notations

Bold uppercase letters ($\mathbf{X}$)	Matrices
Bold lowercase ($\mathbf{x}$)	Vectors
Script lowercase letters (x)	Scalar variables
Calligraphic letters ($\mathcal{X}$)	Sets
μ	A learner that maps the sample space $\mathcal{X} \subset \mathbb{R}^p$ to the sample space $\mathcal{Y} \subset \mathbb{R}$
$\hat{\mu}$	Estimator of μ, $\hat{\mu} \in \mathcal{F}$
$\mathcal{F}$	Class of estimators
#	Cardinality of one set
$[\![n]\!]$	Set {1, ..., n}
$\langle ., . \rangle$	Dot product
π	The shuffling/permutation process
n	Number of observations
p	Number of variables
$\mathcal{S} = \{\mathcal{G}^k, k \in [\![K]\!]\}$	K pre-defined subset of variables or groups
$Tree$	One tree in the Random Forest
N_{Tree}	Number of trees in the Random Forest
i.i.d.	Independent and identically distributed

Synthèse en français

L'accumulation de données au cours des dernières décennies dans divers domaines, notamment l'imagerie médicale, la surveillance de la santé publique et la génomique, a entraîné une augmentation exponentielle du nombre de variables fortement interconnectées, ce qui a rendu difficile le traitement efficace des interrelations entre différents prédicteurs par de simples modèles linéaires ou arborescents. Ces modèles, qui fournissaient auparavant des explications transparentes sur leurs résultats, ont été largement utilisés dans les premiers temps de la science des données. Cependant, ils s'efforcent aujourd'hui de faire face à la complexité accrue des données. Avec l'émergence de l'apprentissage automatique (ML), de nouveaux modèles non linéaires à haute capacité ont été introduits, tels que les forêts aléatoires (RF) et les réseaux de neurones profonds (DNN). Ces méthodes se caractérisent par un accès limité et/ou ininterprétable à leurs paramètres, ce qui leur vaut d'être perçues comme des "boîtes noires" du point de vue humain. Il est donc nécessaire d'élaborer des méthodes d'explication spécifiques pour comprendre leur comportement. Si les méthodes basées sur les permutations sont largement utilisées dans les environnements à faible dimension en raison de leur efficacité de calcul et de leurs garanties statistiques, elles risquent d'interpréter à tort des prédicteurs non pertinents comme étant pertinents dans des environnements à forte corrélation. En outre, dans les cas à haute dimension tels que la neuro-imagerie avec une multitude de régions du cerveau, l'itération sur tous les régions pour évaluer leur degré de signification est une tâche coûteuse.

Pour remédier aux limites de *haute-dimensionnalité* et de *haute-corrélation*, nous avons développé un cadre statistiquement valide pour l'importance des variables, qui peut être appliqué à des cohortes biomédicales à grande échelle. Cette thèse se compose de deux parties : la première partie fournit un contexte et une justification essentiels pour la deuxième partie, qui présente les nouveaux cadres développés.

Dans le chapitre 1, les défis rencontrés dans les environnements à haute dimension et la nécessité de modèles d'apprentissage automatique à haute capacité et non interprétables, communément appelés "boîtes noires", sont examinés. Ensuite, le concept d'importance des variables au niveau individuel et au niveau du groupe est introduit, ce qui est essentiel pour améliorer la transparence de ces modèles. Bien que l'inférence marginale soit largement utilisée, nous discutons de la nécessité de recourir plutôt à l'inférence conditionnelle, car le fait d'ignorer les dépendances entre les prédicteurs conduit à déclarer à tort les prédicteurs comme significatifs. Bien que de nombreuses méthodes d'explication aient été introduites, le maintien d'un con-

trôle statistique sur l'importance des variables reste le facteur clé pour instiller la confiance dans les méthodes d'IA afin que les praticiens puissent comprendre les enseignements tirés des modèles d'apprentissage automatique. Il se peut que ce besoin n'ait pas été considéré comme prioritaire dans les travaux précédents. Dans le chapitre 2, nous examinons la littérature afin d'identifier les méthodes de pointe pour déterminer l'importance des variables, en considérant différentes perspectives (*local* vs *global*, modèle-*spécifique* vs *agnostique*), et en déterminant si des garanties statistiques sont fournies. Pour les environnements à haute dimension, chap. 3 répond au besoin de méthodes basées sur les groupes, où les variables sont regroupées par *data-driven* ou *knowledge-driven* a priori. Enfin, le chapitre 4 présente diverses modalités de données au niveau de la population au sein de grandes cohortes biomédicales, allant de l'imagerie cérébrale aux données sociodémographiques. Ces modalités sont utilisées pour obtenir des informations dans le domaine des neurosciences.

La partie contribution de cette thèse s'articule autour de trois axes principaux, qui sont présentés comme suit.

1 . Les Schémas Conditionnels, le Remède aux *Fausses* Variables Significatives

À l'heure actuelle, l'approche de l'évaluation de l'importance fondée sur la suppression est la méthodologie la plus largement acceptée, en particulier lorsque des garanties statistiques sont recherchées pour justifier l'inclusion de variables. Cette approche est souvent mise en œuvre à l'aide de schémas de permutation variables. Toutefois, ces approches risquent d'identifier à tort des variables sans importance comme étant importantes en présence de corrélations entre les covariables. Dans le chapitre 5, nous présentons une approche systématique pour l'étude de l'importance de la permutation conditionnelle (CPI) qui est agnostique et calculatoirement légère, ainsi que des repères réutilisables d'estimateurs d'importance variable de pointe. Nous démontrons théoriquement et empiriquement que *CPI* surmonte les limites de l'importance de la permutation standard en fournissant un contrôle précis de l'erreur de type I. Lorsqu'il est utilisé avec un réseau de neurones profond, *CPI* fait preuve d'une grande précision sur l'ensemble des critères de référence. Une expérience d'analyse de données réelles dans un ensemble de données médicales à grande échelle a démontré que *CPI* permet une sélection plus parcimonieuse des variables statistiquement significatives. Nos résultats indiquent que *CPI* peut être facilement utilisé pour remplacer les méthodes basées sur la permutation.

Travail publié Chamma, A., Engemann, D., & Thirion, B.. (2023). Statistically Valid Variable Importance Assessment through Conditional Permutations. *In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. DOI: 10.48550/arXiv.2309.07593..

2. Le Regroupement est la Clé des Paramètres à *Haute-Dimension*

Comme indiqué précédemment, l'importance de la permutation conditionnelle (*CPI*) contourne les limites de l'importance de la permutation (PI) dans les cas de forte corrélation. Toutefois, dans des contextes à haute dimension où les corrélations élevées entre les variables annulent leur importance conditionnelle, l'utilisation de *CPI*, ainsi que d'autres méthodes, aboutit à des résultats peu fiables et à des coûts de calcul exorbitants. Le regroupement des variables par le biais d'une classification ou d'une connaissance préalable offre un certain degré de résilience et facilite des interprétations plus robustes. Dans le chapitre 7, nous présentons *BCPI* (importance de permutation conditionnelle basée sur les blocs), un nouveau cadre générique pour le calcul de l'importance variable avec des garanties statistiques qui peuvent être appliquées à la fois aux cas individuels et collectifs. En outre, le traitement de groupes à cardinalité élevée (comme un ensemble d'observations d'une modalité donnée) est à la fois long et gourmand en ressources. C'est pourquoi nous introduisons également une nouvelle approche d'empilement étendant l'architecture DNN avec des couches sous-linéaires adaptées à la structure du groupe. Nos résultats démontrent que l'approche qui en découle, étendue à l'empilement, contrôle l'erreur de type I, même avec des groupes fortement corrélés, et qu'elle présente une précision optimale dans tous les points de référence. En outre, une analyse de données réelles est effectuée sur un ensemble de données médicales à grande échelle dans le but de démontrer la cohérence entre nos résultats et ceux rapportés dans la littérature pour la prédiction d'une sortie biomédicale.

Travail Publié Chamma, A., Thirion, B., & Engemann, D.. (2024). A Variable Importance in High- Dimensional Settings Requires Grouping. *In Proceedings of the 38th Conference of the Association for the Advancement of Artificial Intelligence (AAAI 2024)*. DOI: 10.48550/arXiv.2312.10858..

3 . Importance des Variables pour les Applications de Neuroimagerie

Dans le chap. 8, nous utilisons le cadre construit pour l'importance des variables, BCPI (Cadre construit pour l'importance des variables avec des garanties statistiques), sur des entrées corrélées et des groupes entiers de variables (tels que les lots, les régions cérébrales et les bandes de fréquence) pour déduire quels prédicteurs sont importants. Nous étudions son potentiel pour améliorer les modèles de prédiction d'apprentissage automatique adaptés aux applications neuroscientifiques dans quatre grands ensembles de données multimodales, dans le but de répondre aux questions persistantes concernant l'influence des prédicteurs dans ce domaine. Les résultats de notre étude sont cohérents avec ceux des recherches antérieures sur l'âge du cerveau, tout en fournissant des indications sur l'importance statistiquement valable des prédicteurs individuels.

En Préparation Chamma, A., Engemann, D., & Thirion, B.. (2024). Conditional Permutation Algorithms for Interpretable Machine Learning in Neuroimaging. *To be submitted*

4 . Conclusion

Enfin, dans le chapitre 9, nous concluons la thèse en résumant nos contributions et en offrant des perspectives plus larges sur les questions non résolues qui méritent d'être approfondies.

5 . Autres Travaux

Dans le contexte de l'importance des variables, il y a un manque de données de base pour les variables significatives et non significatives, ce qui nécessite des simulations supplémentaires. Pour garantir le réalisme de ces simulations, deux facteurs clés sont primordiaux : (1) l'inclusion de divers types de variables (continues, ordinales et nominales) et (2) l'incorporation de dépendances entre les différentes variables. Dans le chapitre .6, nous présentons un nouveau cadre itératif pour l'échantillonnage de variables corrélées continues, ordinales et nominales dans un ordre unique. Nous démontrons que la corrélation entre les différentes variables est préservée lors de l'utilisation de variables extraites d'un ensemble de données biomédicales réelles.

6 . Logiciel

Afin de faciliter la reproductibilité scientifique, nous avons implémenté les méthodologies présentées dans cette thèse dans un paquetage Python convivial et interopérable au sein d'un logiciel open-source, qui est disponible en téléchargement sur: https://github.com/Parietal-INRIA/hidimstat. En outre, nous avons préparé une version R de ces méthodologies.

Overview

The accumulation of data over the past decades in various fields, including medical imaging, public health surveillance, and genomics, has led to an exponential increase in the number of highly interconnected variables, which made it challenging for simple linear or tree-based models to effectively process the interrelationships between different predictors. These models, which previously provided transparent explanations of their output, have been used extensively in the early days of data science. However, they are now struggling to cope with the heightened complexity of the data. With the emergence machine learning (ML), new high-capacity non-linear models have been introduced, such as random forests (RF) and deep neural networks (DNN). These methods are characterized by a limited and/or non-interpretable access to their parameters, which has led to them being perceived as "black boxes" from the human perspective. As a result, it is necessary to build dedicated explanation methods to understand their behaviour. While permutation-based methods are widely used in low-dimensional settings due to their computational efficiency and statistical guarantees, they risk misinterpreting non-relevant predictors as relevant in high-correlated settings. Furthermore, in high-dimensional cases such as neuroimaging with a multitude of brain regions, iterating over all regions to gauge their degree of significance is a costly task.

To address both *high-dimensionality* and *high-correlation* limitations, we have developed a statistically valid framework for variable importance that can be applied to large-scale biomedical cohorts. This thesis is comprised of two parts: the first part provides essential background and rationale for the second part, which presents the new developed frameworks.

In chap. 1, the challenges encountered in high-dimensional settings and the necessity for high-capacity, non-interpretable machine learning models, commonly referred to as "black boxes," are discussed. Subsequently, the concept of variable importance at both the single and group levels is introduced, which is essential for improving the transparency of these models. While marginal inference is widely used, we discuss the need to deploy conditional inference instead, as ignoring dependencies among predictors leads to wrongly reporting predictors as significant. While numerous explanation methods have been introduced, maintaining statistical control on variable importance remains the key factor to instill confidence in AI methods for practitioners to comprehend the insights derived from machine learning models. This need might not have been prioritized in previous works. In chap. 2, we examine the literature to identify the state-of-the-art methods for determining variable importance, considering different perspectives (*local* vs *global*, model-

specific vs *agnostic*), and whether statistical guarantees are provided. For high-dimensional settings, chap. 3 addresses the need for group-based methods, where variables are grouped by either *data-driven* or *knowledge-driven* prior grouping. Finally, chap. 4 introduces various population-level data modalities within large biomedical cohorts, ranging from brain imaging to socio-demographics. These modalities are used to extract insights in neuroscience.

The contribution part of this thesis is organized around three major directions, which are presented as follows.

1 . Conditional Schemes, The Remedy for *Fake* Significant Variables

At present, the removal-based approach to importance assessment is the most widely accepted methodology, particularly when statistical guarantees are sought to justify variable inclusion. This approach is often implemented with variable permutation schemes. However, these approaches risk misidentifying unimportant variables as important in the presence of correlations among covariates. In chap. 5, we present a systematic approach for studying conditional permutation importance (CPI) that is model agnostic and computationally lean, as well as reusable benchmarks of state-of-the-art variable importance estimators. We demonstrate theoretically and empirically that *CPI* overcomes the limitations of standard permutation importance by providing accurate type-I error control. When used with a deep neural network, *CPI* consistently shows top accuracy across benchmarks. An experiment on real-world data analysis in a large-scale medical dataset demonstrated that *CPI* provides a more parsimonious selection of statistically significant variables. Our results indicate that *CPI* can be readily used as drop-in replacement for permutation-based methods.

Published work Chamma, A., Engemann, D., & Thirion, B.. (2023). Statistically Valid Variable Importance Assessment through Conditional Permutations. *In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. DOI: 10.48550/arXiv.2309.07593..

2 . Grouping is The Key for *High-Dimensional* Settings

As previously stated, conditional permutation importance (*CPI*) circumvents the limitations of permutation importance (PI) in high-correlation cases. However, in high-dimensional settings where high correlations between variables negate their conditional importance, the use of *CPI*, as well as other methods, results in unreliable outcomes and exorbitant computational costs. The grouping of variables via clustering or prior knowledge provides a degree of resilience and facilitates more robust interpretations. In chap. 7, we introduce *BCPI* (block-based conditional permutation importance), a new generic framework for variable importance computation with statistical guarantees that can be applied to both single and group cases. Furthermore, the handling of groups with high cardinality (such as a set of observations of a given modality) is both time-consuming and resource-intensive. Therefore, we also introduce a new stacking approach extending the DNN architecture with sublinear layers adapted to the group structure. Our results demonstrate that the ensuing approach extended with stacking controls the type-I error even with highly-correlated groups and shows top accuracy across benchmarks. Furthermore, a real-world data analysis is conducted on a large-scale medical dataset with the objective of demonstrating the consistency between our results and those reported in the literature for a biomarker prediction.

Published work Chamma, A., Thirion, B., & Engemann, D.. (2024). A Variable Importance in High- Dimensional Settings Requires Grouping. *In Proceedings of the 38th Conference of the Association for the Advancement of Artificial Intelligence (AAAI 2024)*. DOI: 10.48550/arXiv.2312.10858..

3 . Variable Importance for Neuroimaging Applications

In chap. 8, we utilize the constructed framework for variable importance, BCPI (Built Framework for Variable Importance with Statistical Guarantees), on correlated inputs and entire groups of variables (such as batches, brain regions, and frequency bands) to infer what predictors are important. We investigate its potential to enhance machine learning prediction models tailored for neuroscience applications within four big multimodal datasets, aiming to address lingering questions regarding the influence of predictors in this domain. The findings of our study are consistent with those of previous research on brain age, while also providing insights into the statistically valid importance of individual predictors.

In Preparation Chamma, A., Engemann, D., & Thirion, B.. (2024). Conditional Permutation Algorithms for Interpretable Machine Learning in Neuroimaging. *To be submitted*

4 . Conclusion

Finally, in chap. 9, we wrap up the thesis by summarizing our contributions and offering broader perspectives on unresolved questions that deserve further investigation.

5 . Other Works

In the context of variable importance, there is a lack of ground truth data for significant and non-significant variables, necessitating additional simulations. To ensure the realism of these simulations, two key factors are paramount: (1) the inclusion of diverse types of variables (continuous, ordinal, and nominal) and (2) the incorporation of dependencies among the different variables. In chap .6, we present a novel iterative framework for the sampling of continuous, ordinal, and nominal correlated variables in a single order. We demonstrate that the correlation among the different variables is preserved when using variables extracted from a real biomedical dataset.

6 . Software

In order to facilitate scientific reproducibility, we have implemented the methodologies presented in this thesis within a user-friendly and interoperable Python package within an open-source software, which is available for download on: `https://github.com/Parietal-INRIA/hidimstat`. Additionally, we have prepared an R version of these methodologies.

Part I

Background

1 - Motivation and Problem Statement

Summary The use of Machine Learning (ML) techniques in various scientific fields is due to their ability to predict outcomes based on complex input data. However, for the purpose of advancing scientific understanding, ML should not be used exclusively as a prescriptive tool. Instead, it should be leveraged to generate scientific explanations. This necessitates the incorporation of explainable AI (XAI) methodologies, particularly variable importance (VI) analysis. Variable importance has evolved into an essential instrument, furnishing valuable perspectives for selecting features, interpreting models, and debugging them. By identifying the most relevant variables, researchers can simplify the modeling process and potentially improve performance. Understanding how different variables contribute to a model's predictions is crucial for explaining its logic and building trust in complex AI systems. Furthermore, metrics that gauge variable importance can help identify irrelevant or redundant variables, thus preventing overfitting. In this chapter, we provide a thorough examination of the motivation behind our research, articulate the problem statement, and introduce the foundational concepts utilized in the following chapters.

1.1 . Machine Learning & High-Dimensional Settings

Machine learning (ML) algorithms are widely used in various scientific fields, including biomedical applications [Strzelecki and Badura, 2022, Alber et al., 2019], neuroscience [Kora et al., 2021, Knutson and Pan, 2020], and social sciences [Lundberg et al., 2022, Chen et al., 2021]. The use of machine learning is of growing interest in biomedical research [Iniesta et al., 2016, Taylor and Tibshirani, 2015, Malley et al., 2011] for predicting biomedical outcomes from heterogeneous inputs [Giorgio et al., 2022, Sechidis et al., 2021, Hung et al., 2020]. Biomarker development is increasingly focusing on multimodal data including brain images, genetics, biological specimens, and behavioral data [Yang et al., 2022, Coravos et al., 2019, Castillo-Barnes et al., 2018, Siebert, 2011, Ye et al., 2008]. The growing significance of machine learning in society has raised concerns about accountability, which has led to research on interpretable machine learning.

Indeed, the field of health and life sciences is undergoing exponential growth in both the volume and complexity of data. The analysis of this data, often referred to as high-dimensional, presents a complex and multifaceted challenge for researchers and practitioners. It encompasses diverse formats and emerges from various contexts, requiring advanced analytical methods

for examination. For instance, comprehending treatment response in cancer patients requires navigating a complex network of interconnected variables. These variables include patient characteristics such as age, genetic makeup, and lifestyle factors, disease characteristics such as subtype and stage, treatment protocols such as dosage and frequency, biomarkers such as hormone receptor status and genetic mutations, adverse effects, supportive care measures, and methods for response assessment such as imaging scans and blood tests. When evaluating the effectiveness of a new chemotherapy drug for breast cancer, researchers must consider how these variables interact to determine which patients are most likely to benefit from the treatment, what factors contribute to treatment success or failure, and how to optimize treatment strategies for improved outcomes. Large cohort studies, such as repositories like the UK Biobank [Sudlow et al., 2015], generate vast amounts of complex data due to the large volume and heterogeneity of the variables captured. The data encompasses a rich array of diverse modalities, from discrete binary or categorical indicators to continuous variables that reflect various physiological parameters.

This inherent complexity is further amplified by the high dimensionality of the data, especially in domains like brain imaging and genomic analysis. These fields involve a multitude of variables, posing significant challenges for researchers attempting to extract meaningful insights from the data. For example, a researcher investigating heart disease may first examine patient demographics and medical history. By utilizing more advanced methods, researchers may discover subtle genetic variations or lifestyle habits that have a significant impact on heart health. This could lead to the development of more targeted prevention strategies. In these domains, the data not only increases in scale but also forms a complex web of interrelationships. The variables' intricate relationships result in highly correlated features and patterns that are challenging to detect. To comprehend these relationships, researchers require powerful analytical tools. High-dimensional settings with correlated inputs can put strong pressure on model identification.

Accordingly, when dealing with large samples where the number of observations greatly exceeds the number of variables, it is of interest to use complex prediction models [Biecek, 2018]. While complex, often high-capacity nonlinear models, offer greater predictive power compared to simpler linear or tree-based models, they concurrently present challenges in obtaining relevant explanations [Mi et al., 2021], thus it becomes harder to assess the role of features in the prediction [Casalicchio et al., 2019, Altmann et al., 2010]. Additionally, in epidemiological and clinical studies, one is interested in *population-level* feature importance, as opposed to *instance-level* feature importance. There's a fundamental contrast between the two approaches: local methods seek explanations specific to individual instances or samples,

whereas global methods seek overarching explanations regarding the underlying mechanism generating the observed data.

Medical research seeks to identify a limited number of crucial factors for diagnosis or risk assessment by utilizing large observational datasets and robust predictive models. This approach aids in efficiently targeting populations and reducing the costs associated with extensive phenotyping in medical studies. Similarly, in studies focusing on treatment response prediction for a given disease [Hines et al., 2023], the emphasis is placed on deriving predictions from a restricted set of variables, enhancing practical applicability.

1.2 . Single Variable Importance

When a patient presents with symptoms such as fever, cough, and fatigue, it is crucial to identify the relevant variables for their condition. For instance, the presence of fever may indicate a possible infection or inflammatory response, cough may indicate respiratory problems like pneumonia, and fatigue may indicate various underlying conditions. This information is critical for formulating an accurate diagnosis and treatment plan. Failure to do so could result in a misdiagnosis, leading to ineffective or even harmful interventions. Consider also a scenario where researchers analyze a massive dataset of brain scans, cognitive tests, sleep patterns, and social interactions from thousands of people to determine the factors that contribute to a mental health disease like depression. In both cases, the challenge is to distinguish the truly influential factors from the available information. This is where *Variable Importance* comes into play.

According to Zien et al. [2009], variable importance (VI) can be introduced as the process of "estimating the influence of a given input variable to the prediction made by a model", thus reaching a comprehensive understanding of the decision process which is crucial for providing statistical and, ideally, scientific insights to the practitioner [Gao et al., 2022, Molnar et al., 2021a, Fleming, 2020, Hooker et al., 2019].

The concept of variable importance has a rich historical background that is closely linked to the development of statistical modeling, despite its apparent simplicity. Since the early days of scientific investigation, researchers have aimed to identify the significance of specific factors through isolation in experiments [Hepburn and Andersen, 2021]. With the rise of statistical modeling, particularly linear regression [Galton, 1886], a more quantitative approach for evaluating the importance of variables has been developed. The magnitude and significance of a variable's coefficient offered a measurable gauge of its impact. Methods such as stepwise regression were developed to systematically identify the most influential variables by adding or removing them from the model. Nonetheless, these methods focus primarily on individual vari-

ables and linear relationships, which can make it challenging to identify complex interactions between variables that may be crucial for understanding the underlying connections in the data. Therefore, more advanced techniques are needed. These advanced techniques are necessary for uncovering complex patterns and relationships among the variables.

A significant moment occurred in the 1990s with the introduction of specific metrics designed for tree-based models, such as Random Forests. In his influential paper, Breiman [2001] introduced the Mean Decrease Accuracy (MDA) metric. This metric evaluated the reduction in a model's accuracy when the values of a variable were randomly permuted, effectively quantifying the variable's importance in accurate prediction. This progress led to the development of various metrics for evaluating the significance of variables, which will be detailed in the next chapter. It is relevant to note the difference between analyzing a model, which involves identifying variables that contribute to a specific outcome, and analyzing the process that produces the data. This study aims to identify relevant predictors for the model by using non-causal techniques based on the statistical relationships observed between the variables. However, the quality of interpretation heavily relies on the model used, emphasizing the importance of selecting an appropriate fitting model. This highlights the rationale for using model-agnostic approaches, which give scientists the chance to explain the decision-making process of any model.

Several model-agnostic methods have been proposed [Molnar, 2022, Ribeiro et al., 2016]. Examples include *Permutation Feature Importance (PFI)* [Breiman, 2001], *Conditional Randomization Test* [Candes et al., 2017] and *Leave-One-Covariate-Out (LOCO)* [Lei et al., 2018]. All these instances constitute removal-based approaches [Covert et al., 2020], and are so far, the only ones known to provide statistically sound measures of significance. The need for statistical significance in hypothesis testing is discussed in the following sections. Importantly, removal-based approaches require retraining the model after removing the variable of interest and are, therefore, time-consuming. Moreover, the common Permutation Importance (*PI*, Breiman 2001) risks mistaking insignificant variables for significant ones when variables are correlated [Hooker et al., 2021].

1.3 . Hypothesis Testing

Almost a century ago, Ronald Fisher introduced the concept of hypothesis testing in his seminal works [Fisher, 1992, 1936]. Hypothesis testing involves using observed data to make decisions regarding the characteristics of the underlying data-generating model. A hypothesis test sets out rules specifying for which sample values the decision is made to accept $\mathcal{H}_0$, the null hypothesis, and for which sample values $\mathcal{H}_0$ is rejected and $\mathcal{H}_\alpha$, the alternative hypothesis, is accepted.

An early instance of hypothesis testing is recounted in Fisher [1936]'s book "Design of Experiments". The renowned statistician sought to evaluate a female colleague's assertion that she could discern whether milk or tea was poured first into a cup based solely on taste. Fisher devised an experiment in which his colleague was presented with eight cups of tea, four prepared with milk first and four with tea first, in random sequence. Subsequently, one might inquire about the probability of her correctly identifying the origins of the beverages purely by chance. The null hypothesis, denoted as $\mathcal{H}_0$, posited that the lady lacked the ability to distinguish between the order of preparing a tea cup, while the alternative hypothesis, $\mathcal{H}_\alpha$, suggested that she could accurately classify the order of tea cup preparation.

Usually, in conducting hypothesis testing, we define a test statistic, indicated by T. An example of a test statistic is the sample mean: $T(x_1, \ldots, x_n)$ = $\sum_{i=1}^{n} \frac{1}{n} x_i$. Once the test statistic is computed, a decision must be made to determine whether the outcome of the hypothesis test holds statistical significance. One approach involves comparing the test statistic to a specific quantile of its null distribution, known as the significance level (denoted by α). This significance level represents the probability of rejecting the null hypothesis $\mathcal{H}_0$ under the assumption that it is true. Alternatively, one may present the outcome of a hypothesis test using a p_{value} which satisfies under $\mathcal{H}_0$ the following property: $\mathbb{P}_{\mathcal{H}_0}(p_{value} \leq t) \leq t \ \forall t \in [0, 1]$. As a consequence, it is feasible to create a test statistic at the significance level α, for any $\alpha \in (0, 1)$. Essentially, the smaller the p_{value}, the greater the confidence with which the statistician can reject $\mathcal{H}_0$.

In any testing scenario, there are two potential errors that can occur. A false positive, also known as a type I error, happens when we incorrectly reject a true null hypothesis. Conversely, a false negative, or type II error, occurs when we fail to reject an alternative hypothesis. Ideally, one aims to minimize both the occurrences of type I and type II errors simultaneously. However, achieving this balance is often impractical, necessitating a trade-off between the two. Typically, this trade-off entails minimizing type II errors while adhering to a constraint on type I errors. Minimizing false negatives can also be interpreted as maximizing true discoveries, which represents the statistical power of the hypothesis test. The statistical power of an individual hypothesis

test refers to the likelihood that the test accurately rejects the null hypothesis $\mathcal{H}_0$ when the alternative hypothesis $\mathcal{H}_\alpha$ is true.

1.4 . Statistical Control

In the latter part of the 20th century, there was a surge in the development and deployment of sophisticated machine learning models. The field of interpretable machine learning focuses on obtaining statistical guarantees for variable importance under these models. The goal is to identify the key variables that drive a model's predictions using mathematically proven methods. These guarantees are a crucial in linking a model's complexity to human comprehension, which enables trust in the model's interpretations. They ensure that identified important variables are not merely bystanders or victims of spurious correlations. Moreover, these guarantees facilitate model optimization by discarding irrelevant features, enhancing efficiency.

One aspect often neglected in much of the literature on variable importance is the crucial need to manage error rates during the process. The potential costs of errors can be large; for instance, in genetic analysis, the expense of investigating a wrongly identified gene could be unbearable [Zhao et al., 2022, Candes et al., 2017]. As the identification of relevant variables is model-dependent and potentially unstable, point estimates of variable importance are misleading. One needs confidence intervals of importance estimates or statistical guarantees, such as type-I error control, i.e. the percentage of non-relevant variables detected as relevant (false positives). This control depends on the accuracy of the p-values on variable importance being non-zero [Cribbie, 2000].

Within the family of removal-based importance assessment methods [Covert et al., 2022], a popular model-agnostic approach is *permutation* variable importance, that measures the impact of shuffling a given variable on the prediction [Janitza et al., 2018, Breiman, 2001]. By repeating the *permutation* importance analysis on permuted replicas of the variable of interest, importance values can be tested against the null hypothesis of being zero, yielding p-values that are valid under interactions among variables. Yet, statistical guarantees for permutation importance assessment do not hold in the presence of correlated variables, leading to selection of unimportant variables [Molnar et al., 2021b, Hooker et al., 2021, Nicodemus et al., 2010]. For instance, the method proposed in [Mi et al., 2021] is a powerful variable importance evaluation scheme, but it does not control the rate of type-I error.

As models play an increasingly central role in critical decision-making, the need for trust and interpretability becomes more important. Statistical guarantees can increase trust in AI techniques and even democratize interpretability, making it possible for wider audiences to understand the insights

obtained from machine learning models. Researchers have developed new approaches, such as Conditional Permutation Importance (CPI) [Watson and Wright, 2021, Debeer and Strobl, 2020], to address these challenges, aiming for robust guarantees that apply to all model types. These guarantees represent a significant step towards responsible and trustworthy AI, transforming machine learning models from powerful black boxes into transparent partners in unraveling the world's complexities.

1.5 . Conditional Inference

As previously mentioned, the main challenge in analyzing variable importance is accounting for dependencies among variables. In many cases, these dependencies appear as correlations, while the underlying relationships between variables, often represented by a directed acyclic graph, can only be partially inferred from existing knowledge or fundamental principles. It is of interest to assess whether a given measurement is worth acquiring, on top of others, for a diagnostic or prognostic task. A conventional differentiation, as outlined in the work by Candes et al. [2017], must be made between the easily accessible methods assessing *marginal feature importance* and *conditional feature importance*.

Marginal feature importance assesses the marginal relationship between a single variable x^j and the response y, without considering interactions with other variables $\mathbf{x}^{-\mathbf{j}}$. However, as datasets become more complex with advancements in data acquisition technologies, it is increasingly apparent that focusing solely on individual variables is inadequate [Smith and Nichols, 2018]. Instead, attention must be directed towards uncovering and comprehending the nuanced relationships and dependencies that emerge through interactions between variables [Akogul, 2023]. Ignoring these interactions risks oversimplification and can lead to incomplete or even misleading conclusions [Strobl et al., 2008]. For example, in the case of fMRI scans, it is widely acknowledged that neighboring brain voxels exhibit positive correlations. During cognitive activities such as watching a video, activation usually occurs. observed across multiple brain voxels simultaneously. Hence, it is more relevant to investigate the conditional relationship between y and x^j: When watching a video, does the activation of voxel x^j in the brain predict a specific feature in the video, taking into account its interactions with other brain voxels $\mathbf{x}^{-\mathbf{j}}$? This concept is known as *conditional feature importance* or *conditional inference*. Fig. 1.1 provides an illustration of this concept within the context of neuroscience.

Conditional inference builds upon the concept of conditional probability, which was first introduced in the 17th century as part of the early developments in probability theory [Bayes and Price, 1763]. Significant contribu-

tions to statistical theory and methodology were made in the early 20th century [Neyman and Pearson, 1933], particularly in areas such as experimental design, which further strengthened the foundation for conditional inference. However, the widespread adoption and formalization of conditional inference truly flourished in the latter half of the 20th century [Stigler, 1999, Nelder and Wedderburn, 1972]. This coincided with significant advancements in computing power and the development of more sophisticated statistical methodologies. Conditional inference became a powerful tool for addressing the challenge of controlling for confounding variables to extract valid conclusions from observational and experimental data [MA and JM, 2020]. In fact, the recent decades have witnessed a surge in the application of conditional inference across diverse fields [Liu et al., 2012, Chatterjee and Carroll, 2005]. This includes epidemiology, economics, and the social sciences, where researchers deal with intricate datasets. These datasets often necessitate sophisticated analytical techniques to disentangle causal relationships amidst a web of influencing factors. *Conditional feature importance* can overcome the limitations of *marginal importance* [Chamma et al., 2023, Blesch et al., 2023, Watson and Wright, 2021, Debeer and Strobl, 2020, Fisher et al., 2019]. It continues to evolve alongside advancements in statistical theory and computational methods.

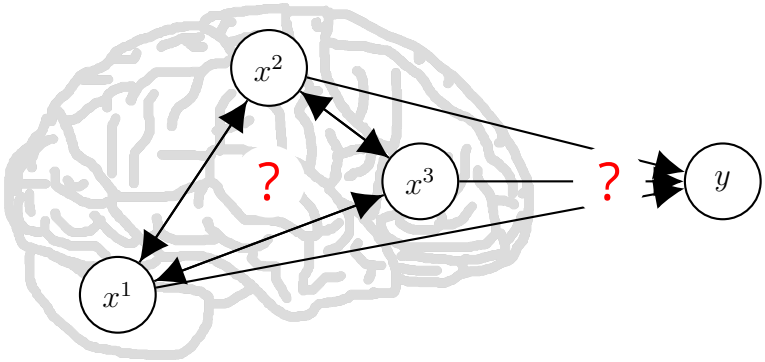

Figure 1.1: **Conditional Inference in Neuroscience**: Visualization of the interplay among three brain voxels within cognitive neuroscience and a behavioral variable y, usually induced by a controlled stimulus. The objective is to explore whether the activation in one voxel causes a certain mental task, taking into account its interaction with the other two brain voxels. Figure is derived from [Weichwald and Peters, 2020, Weichwald et al., 2015].

1.6 . Group Variable Importance & Stacking

Although the field of machine learning has a long history, the idea of group variable importance is a relatively recent development. In high-dimensional settings, single variable importance computation suffers from very high correlation between the variables [Chevalier et al., 2021]. More precisely, this makes conditional importance estimation less informative, as it remains unclear how much information each variable adds. In the extreme case where variables are duplicated, conditional importance can no longer be defined. More generally, correlations larger than $.8$ are known to present a hard challenge, at least for linear learners [Chevalier et al., 2021]. Importance analysis then typically yields spuriously significant variables, which ruins its ability to statistically control the false positive rate [Strobl et al., 2008]. As previously mentioned, the huge increase in data volume over the past decades has led to datasets becoming significantly larger and more complex, with numerous variables. Examining each of the hundreds or thousands of variables separately would result in prohibitively high computation costs [Covert et al., 2020] —removal procedures typically have cubic complexity due to the refitting process— and defy model interpretability.

Group variable importance provides a solution by evaluating the importance of clusters of variables. For instance, in a study that investigates the neural correlates of depression severity in adolescents, researchers use group variable importance analysis to clarify the relative contributions of various brain regions to depressive symptomatology. Neuroimaging variables are categorized into three main groups: structural measures, functional connectivity patterns, and task-based activation patterns. Structural measures include gray matter volume and cortical thickness. Functional connectivity patterns refer to resting-state functional connectivity between key brain regions. Task-based activation patterns involve activation levels during emotion processing tasks. By concentrating on groups, a more comprehensive understanding of how variables interact and collectively influence the model's performance can be achieved. Group-based analysis can regularize power estimates and lead to reduced computation time [Molnar et al., 2021b, Bühlmann, 2013]. This can improve inference as it helps handle the curse of correlated variables in high-dimensional settings. So far, common group-based methods have neglected investigating statistical guarantees, in particular, type-I error control, i.e. the percentage of irrelevant variables identified as relevant (false positives). Statistical error control for groups obviously requires information on variable grouping available through two strategies: *Knowledge-driven* grouping, where the variables are grouped based on their domain-specific information rather than their shared statistical properties and *Data-driven* grouping, where clustering approaches are used such as hierarchical or divisive clustering.

Grouping has also been successfully performed for multimodal learning

problems [Albu et al., 2023, Engemann et al., 2020, Rahim et al., 2015] via model stacking [Wolpert, 1992] which is typically based on pipelines of disconnected models. The variables from each group are combined into stacked features, integrating diverse sources of information.

1.7 . Problem Setting

1.7.1 . Single Case

We consider the regression or the classification problem where the response vector $\mathbf{y} \in \mathbb{R}^n$ or $\in \{0,1\}^n$ respectively and the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ (encompasses n observations of p variables) where the i^{th} row and the j^{th} column are denoted $\mathbf{x_i}$ and $\mathbf{x^j}$ respectively. Let $\mathbf{X^{-j}} = (\mathbf{x^1}, \ldots, \mathbf{x^{j-1}}, \mathbf{x^{j+1}}, \ldots, \mathbf{x^p})$ be the design matrix, where the j^{th} column is removed, and $\mathbf{X^{(j)}} = (\mathbf{x^1}, \ldots, \mathbf{x^{j-1}}, \{\mathbf{x^j}\}^\pi, \mathbf{x^{j+1}}, \ldots, \mathbf{x^p})$ the design matrix with the j^{th} column shuffled. The rows of $\mathbf{X^{-j}}$ and $\mathbf{X^{(j)}}$ are denoted $\mathbf{x_i^{-j}}$ and $\mathbf{x_i^{(j)}}$ respectively, for i $\in [\![n]\!]$.

1.7.2 . Group Case

Let $\mathcal{J} = \{j_1, \ldots, j_r\}$ be a subset of r variables with consecutive indices in $[\![p]\!]$, $r \leq p$. We indicate the $\mathcal{J}^{th}$ subset of columns by $\mathbf{X}^{\mathcal{J}}$. Let $\mathbf{X}^{-\mathcal{J}} = (\mathbf{x^1}, \ldots, \mathbf{x^{j_1-1}}, \mathbf{x^{j_r+1}}, \ldots, \mathbf{x^p})$ be the design matrix with the $\mathcal{J}^{th}$ subset of variables is removed. Let $\mathbf{X}^{(\mathcal{J})} = (\mathbf{x^1}, \ldots, \mathbf{x^{j_1-1}}, \{\mathbf{x^{j_1}}\}^\pi, \ldots, \{\mathbf{x^{j_r}}\}^\pi, \ldots, \mathbf{x^p})$ be the design matrix with the $\mathcal{J}^{th}$ subset of variables is shuffled. The rows of $\mathbf{X}^{-\mathcal{J}}$ and $\mathbf{X}^{(\mathcal{J})}$ are denoted $\mathbf{x_i}^{-\mathcal{J}}$ and $\mathbf{x_i}^{(\mathcal{J})}$ respectively, for $i \in [\![n]\!]$.

Across this work, we rely on an i.i.d. sampling train/validation/test partition scheme where the n samples are divided into n_{train} training and n_{test} test samples. The train samples were used to train $\hat{\mu}$ with empirical risk minimization. This function is utilized for appraising the importance of variables on a novel dataset (test set).

1.8 . Conclusion

The concept of variable importance, at both single and group levels, has evolved from its rudimentary beginnings to become a fundamental pillar of modern data science. It empowers researchers not only to construct intricate models but also to comprehend their inner workings. By tackling hurdles such as correlated variables and curse of dimensionality, variable importance stands ready to assume an even more significant role in encouraging trust and ethical advancement in the constantly evolving domain of AI. Furthermore, there is interest in exploring conditional independence relationships e.g. among brain regions and cognitive or behavioral outcomes, highlighting the diverse applications of this research approach.

2 - Variable Importance Methods

Summary Analyzing the importance of individual variables is a complex issue within the spectrum of model interpretability. The employed model can be in certain situations a transparent-box, providing simple and interpretable access to its internal architecture and parameters such as the coefficients of the *Linear Models* and the depth of the conditioning feature in the *Decision trees*. In other situations, the chosen model can be seen as a black-box because of its complex architecture and massive number of parameters, thus limiting the understandable access to only the input/output sources, a.k.a. *Sensitivity Analysis*. *Sensitivity analysis* is a set of techniques aimed at quantifying the influence of each variable or group of variables on the model predictions. In addition, there is no one-size-fits-all interpretation method for different tasks. With *Non-perturbative sensitivity analysis* methods, the *VI* is examined directly without introducing any changes to the input/output sources, such as saliency maps. Nevertheless, *Perturbative sensitivity analysis* or *Removal-based* methods evaluate the impact of the single features by the mean of some perturbation introduced to the variable of interest. These methods can be *instance*-based or *population*-based, *specifically* tailored for a deployed model or *agnostic* without any assumptions related to the predictive model, either through a *single fit* or conducting a *refitting* process. In this chapter, we explore the literature to highlight the state-of-the-art methods for assessing single variable importance. Table 2.1 provides a summary of the methods employed in this chapter. These include model-agnostic vs model-specific approaches, removal-based vs non-perturbative approaches, global-based vs instance-based approaches, which provide or do not provide statistical guarantees.

Method	Model-agnostic?	Removal based?	Global?	Statistical guarantees?
Marginal	No	No	Yes	FPR
MDI	No	No	Yes	No
BART	No	No	Yes	No
Saliency	No	No	Yes	No
LIME	Yes	No	No	No
d0CRT	No	Yes	Yes	FPR
HoldOut	No	Yes	Yes	FPR
Permuting y	No	Yes	Yes	FPR
Conditional-RF	No	Yes	Yes	No
DeepPINK	No	Yes	Yes	FDR
ALE	Yes	Yes	No	No
SHAP	Yes	Yes	No	No
SAGE	Yes	Yes	Yes	No
Knockoff	Yes	Yes	Yes	FDR
MDA-Permfit	Yes	Yes	Yes	FPR
LOCO	Yes	Yes	Yes	FPR

Table 2.1: **Summarizing table for single-based methods**: This table provides a summary of the methods presented in this chapter, categorizing them into three main groups: model-agnostic vs model-specific, removal-based vs non-perturbative and global vs instance-based approaches. It indicates whether each method provides statistical guarantees or not. *FDR*: False Discovery Rate. *FPR*: False Positive Rate.

2.1 . Non-perturbative Sensitivity Analysis

2.1.1 . Model Specific Methods

Marginal Importance [Jamshidian et al., 2007]

Because there is a need for a baseline to compare with in dealing with the different methods, we integrate the *Marginal Importance* scores of the different predictors. Each variable $\mathbf{x^j} \in \mathbb{R}^n$ is fitted separately to predict the response $\mathbf{y}$ using a *Linear Regression* model as:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x^j}$$

The coefficient β_1, extracted from the model, reflects the VI score of the variable $\mathbf{x^j}$. The *F-test* ($\mathcal{H}_0 : \beta_1 = 0$ vs $\mathcal{H}_1 : \beta_1 \neq 0$) is used to extract the corresponding p-values. The F-statistic is computed as $F = \frac{\text{Mean Squared Regression}(MSR)}{\text{Mean Squared Error}(MSE)}$ where $MSR = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{1}$ having $\bar{y}$ the mean of the observed values of the outcome y and $MSE = \frac{\sum_i^n (y_i - \hat{y}_i)^2}{n-2}$. Using the com-

puted F-statistic and the degrees of freedom of the numerator (1) and the denominator $(n-2)$ from the *F-Table*, we can extract the corresponding p-value denoted as p^j. This method offers statistical guarantees by controlling the type-I error rate for the significance of individual variables, a.k.a the False Positive Rate (FPR).

Mean Decrease Impurity (MDI)

As outlined in the work of [Sutera et al., 2021, Louppe et al., 2013], the variable of interest's importance is revealed through *MDI* as the impact it has on the impurity function used in building the *RF* model. An impurity function is a metric used in decision tree algorithms to evaluate the quality of a split in the dataset. It measures how effectively a split separates the data into homogeneous groups. Common impurity functions used to create classes based on the target variable include Gini impurity, entropy, and misclassification error. The importance score is computed using:

$$VI^j = \frac{1}{N_{Tree}} \sum_{Tree} \sum_{t \in Tree : v(s_t)=j} p(t) \Delta i(s_t, t)$$

where *p(t)* is the proportion $\frac{N_t}{N}$ of samples reaching the node *t*, *N* is the size of the learning sample, $v(s_t)$ is the variable used in the split s_t and $\Delta i(s_t, t)$ is the decrease in the impurity as:

$$\Delta i(s_t, t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

where *i(t)* is the impurity measure in the corresponding node *t*, $p_L = \frac{N_{t_L}}{N_t}$ and $p_L = \frac{N_{t_R}}{N_t}$. *L* and *R* stands respectively for the Left and Right childs of the corresponding node *t*. This method does not offer statistical guarantees.

Bayesian Additive Regression Trees (*BART*)

Chipman et al. [2010] unveiled *BART* as a method inspired by the boosting algorithms, consisting of a sum-of-trees, where each is constrained by a prior regularization to be a weak learner. The fitting and inference are accomplished via an interactive Bayesian backfitting Markov Chain Monte Carlo (MCMC) algorithm.

A collection of trees is built, with each tree designed to avoid handling high levels individually, as the inclusion of large tree components could overpower the intricate structure of the *sum-of-trees* model. The problem is defined as

$$y_i = \sum_{Tree} g(\mathbf{x_i}; Tree, M_{Tree}) + \epsilon_i \quad \forall i \in [\![n]\!], \epsilon \sim N(0, \sigma^2)$$

where $g(\mathbf{x_i}; Tree, M_{Tree})$ is the function that allocates $m \in M_{Tree}$ to $\mathbf{x_i}$, M_{Tree} is a set of parameter values associated with each of the terminal nodes of tree

$Tree$ and $\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}$. The prior df ν and scale λ are calibrated using a rough data-based overestimate, denoted $\hat{\sigma}$ of σ. There are two options for calculating $\hat{\sigma}$ (1) the *naive* approach, which derives $\hat{\sigma}$ from the sample standard deviation of $\mathbf{y}$, and (2) the *linear model* approach, which derives $\hat{\sigma}$ from the residual standard deviation obtained from a least squares linear regression of $\mathbf{y}$ on the original $\mathbf{X}$ variables. A value of ν between 3 and 10 is selected to shape the distribution. Regarding λ, it is chosen to ensure that the qth quantile of the prior distribution on σ corresponds to $\hat{\sigma}$, guaranteeing that P(σ < $\hat{\sigma}$) = q. Different values of q, such as 0.75, 0.9 or 0.99, are considered to effectively position the distribution below $\hat{\sigma}$.

By keeping track of covariate inclusion frequencies, *BART* can identify which components are more important for explaining $\mathbf{y}$ i.e. the proportion of all splitting rules that utilize the j^{th} component of $\mathbf{X}$. This method does not offer statistical guarantees.

2.1.2 . Model Agnostic Methods

Saliency Maps

A saliency map is a topographic map used in computer vision to represent the saliency of various locations in an image. Itti et al. [1998] introduced the concept of saliency, which refers to how distinct a stimulus is compared to its surroundings. This makes it possible to highlight regions that have a significant impact on the understanding of visual content. The saliency map is computed by combining feature maps that represent intensity, color, and orientation. As an example, a self-driving car that is equipped with a camera uses a convolutional neural network (CNN) to identify pedestrians. Once a person is successfully detected on the crosswalk, a saliency map can be generated to highlight their figure [Simonyan et al., 2014]. This helps engineers determine if the CNN focused on relevant parts, such as the body, rather than distracting elements like a colorful bag. This ensures that the car prioritizes the most important information for safe navigation. This method does not offer statistical guarantees.

Local Interpretable Model-agnostic Explanations (*LIME*)

Interpreting a black-box model can be complex because its internal workings are not easily understandable. To address this issue, Ribeiro et al. [2016] proposed the use of local surrogate models to explain black-box predictions through an interpretable local model. Rather than training a global surrogate model, *LIME* emphasizes training local surrogate models to provide explanations for individual predictions.

The dataset is divided following a train/test scheme. Each observation in the test set is regarded as point of interest.

$$\text{explanation}(\mathbf{x}) = \underset{g \in G}{argmin}\ L(\mu, g, \pi_x) + \Omega(g)$$

where G is the family of local interpretable models (e.g. linear regression model), L is the loss function, π_x specifies the size of the neighborhood surrounding instance of interest $\mathbf{x}$, and $\Omega(g)$ is the model complexity. After selecting the instance of interest, the train set is perturbed, yielding black-box predictions for the modified points. These new samples are weighted based on their proximity to the instance of interest. Subsequently, a weighted interpretable model is trained using the dataset incorporating the variations. The prediction is interpreted by examining the explanations provided by the local model. The result is a (n_{test}, p) matrix. Finally, the average of the explanations per variable is computed. This method does not provide statistical guarantees.

2.2 . Removal-based Methods

Many methods can be subsumed under the general category of removal-based approaches [Covert et al., 2022].

2.2.1 . Model Specific Methods

d0CRT

As defined in the work by Liu et al. [2021], $d0crt$ is a method proposed for fast Conditional Randomization Testing (CRT) [Candes et al., 2017]. With the *CRT* proposed by Candes et al. [2017], the association between the outcome $\mathbf{y}$ and the variable of interest $\mathbf{x^j}$ conditioned on $\mathbf{X^{-j}}$ is estimated. The variable of interest is sampled conditionally on the other covariates multiple times to compute a test statistic and p-values. However, this solution is limited to generalized linear models and is computationally expensive. Thus, distillation serves as an acceleration through the computation of the common parts.

For the distillation of $\mathbf{X}$, the chosen index j is dropped resulting in $\mathbf{X}$ = ($\mathbf{X^{-j}}$, $\mathbf{x^j}$). $\mathbf{x^j}$ is predicted by providing the remaining components of $\mathbf{X}$ as the predictors as shown in eq. 2.1. The residuals of the prediction of $\mathbf{x^j}$ denoted $\mathbf{x^{res_j}}$ are along with the standard deviation of the residuals $\sigma_{\mathbf{x^j}}$ as in 2.2 and 2.3.

$$\mathbf{x^j} = \mathbf{X^{-j}}.\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

$$\mathbf{x^{res_j}} = \mathbf{x^j} - \hat{\mathbf{x}}^{\mathbf{j}} \tag{2.2}$$

$$\sigma_{\mathbf{x^j}} = \frac{1}{n}\sum_{i=1}^{n}(x_i^{res_j})^2 \tag{2.3}$$

For the distillation of $\mathbf{y}$, the chosen index j is dropped again resulting in $\mathbf{X}$ = ($\mathbf{X^{-j}}$, $\mathbf{x^j}$). The outcome vector **y** is predicted by providing the remaining components using the RF model. The residuals of the outcome and the standard deviation of the residuals labeled respectively $\mathbf{y^{res_j}}$ and $\sigma_{\mathbf{y^j}}$ are computed in eqs. 2.2 and 2.3.

For the computation of the test statistic, the resampling-free approach of *d0crt* is applied where

$$VI^j = \frac{\mathbf{y^{res_j}}.\mathbf{x^{res_j}}}{n \times \sigma_{\mathbf{x^j}} \times \sigma_{\mathbf{y^j}}}$$

Finally, for the computation of p-values, it is assumed that the test statistic VI is standard Gaussian. In this case, p-values are computed using the Gaussian distribution bilateral test with the cumulative distribution function of the standard Gaussian distribution F_{norm} as $p^j = 1 - F_{norm}(VI^j)$. This method offers statistical guarantees by controlling the type-I error rate for the significance of individual variables.

HoldOut

Janitza et al. [2018] proposed a new fast variable importance test when dealing with high-dimensional data. They declared that, on one hand, a non-positive importance score is sufficient to highlight the non-relevance of the variable. On the other hand, positive importance score is not a sufficient evidence for the relevance of the variable of interest. Therefore, a testing procedure is required to verify if the variable is truly relevant or the result of some randomness.

First, the training data with n samples is split into 2 non-overlapping parts used separately for building two RF models. The second part is used in a later phase for the computation of the variable importance.

$$\underbrace{O_1, O_2, O_3, \ldots, O_{\frac{n}{2}}}_{\substack{\text{Building } Forest_2 \\ \text{VI for } Forest_1}} \underbrace{O_{\frac{n}{2}+1}, \ldots, O_{n-1}, O_n}_{\substack{\text{Building } Forest_1 \\ \text{VI for } Forest_2}}$$

Thus, in a classification setting, each variable has two variable importance scores, one for each split, computed as:

$$VI^{j,l} = \frac{1}{N_{Tree}} \sum_{Tree=1}^{N_{Tree}} \frac{1}{\#S^l} \sum_{i \in S^l} \{I(y_i \neq \hat{y}_i^{*,Tree}) - I(y_i \neq \hat{y}_i^{Tree})\} \tag{2.4}$$

with S^l is the set of observations not used to build the l^{th} forest and $\hat{y}^{Tree}$ and $\hat{y}^{*,Tree}$ denoting the predictions by the $Tree^{th}$ tree before and after permuting the values of the variable j respectively. The variable importance scores over the RFs per variable are averaged as $VI^j = \frac{1}{2}\sum_{l=1}^{2} VI^{j,l}$. As a result, 3 sets, M_1, M_2 and M_3, are extracted designating the negative, zero and negative multiplied by -1 variable importance scores. Next, the null distribution $\hat{F}_0$ is approximated as the empirical cumulative distribution function of M = $M_1 \cup M_2 \cup M_3$. Finally, the p-value is determined as $p^j = 1 - \hat{F}_0(VI^j)$.

To sum up, they approximate the null distribution based on the observed importance scores to provide p-values. Yet, this coarse estimate of the null

distribution can give unstable results. This method provides statistical guarantees by controlling type-I error rate for the significance of individual variables.

Permuting y

Altmann et al. [2010] proposed an alternative method for the *HoldOut* importance when the reconstruction of the null distribution $\hat{F}_0$ lacks non-positive importance scores. They constructed an original RF model where the importance variable scores of the different predictors were computed as in eq. 2.4 by replacing the set of observations not used to build the RF model (S^l) with the OOB (Out-Of-Bag) samples. These scores are denoted VI^{org}.

Subsequently, the outcome y is randomly permuted. Consequently, a novel RF model is constructed with the predictor variables and the permuted outcome, where the importance scores are calculated in accordance with the aforementioned methodology. This process is repeated B times, resulting in each variable j carrying a set of B variable importance scores, designated as $\mathbf{VI^j}$. Finally, the p-values are computed using the non-parametric approach as: $p^j = \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}_{VI_i^j > VI^{org}}$. This method offers statistical guarantees by controlling the type-I error rate for the significance of individual variables.

Conditional-RF

In their work, Strobl et al. [2008] proved empirically that the importance score method for the *RF* model is biased in favor of the non-relevant variables correlated with the relevant ones. They proposed a new method based on conditioning on the remaining variables (or a subset of these variables) which can reflect the true impact more reliably.

First, in each tree, they compute the Out-of-bag (OOB) prediction accuracy before permuting (bp) the values of the variable of interest with:

$$pred_{bp}^{(Tree)} = \frac{\sum_{i \in B^{(Tree)}} I(y_i = \hat{y}_i^{(Tree)})}{\#B^{(Tree)}}$$

where *Tree* is the corresponding tree, I is the identity function and $B^{(Tree)}$ is the Out-of-bag sample for tree *Tree*.

For all variables *Z* to be conditioned on, they extract the breakpoints that split each of these variables in the current tree (*Tree*) and create a grid by bisecting the sample space in each breakpoint. Within this grid, they permute the values of $\mathbf{x^j}$ and recompute the Out-of-bag (OOB) prediction accuracy after permutation (ap):

$$pred_{ap}^{(Tree,j)} = \frac{\sum_{i \in B^{(Tree)}} I(y_i = \hat{y}_{i,\pi^j|Z}^{(t)})}{\#B^{(Tree)}}$$

where $\hat{y}_{i,\pi^j|Z}^{(Tree)} = \hat{\mu}^{(Tree)}(X_{i,\pi^j|Z})$ is the predicted class for observation *i* after

permuting the values of the variable $\mathbf{x^j}$ within the grid defined by the variables *Z*.

Finally, they compute the variable importance of the variable of interest $\mathbf{x^j}$ in one tree *Tree* as the difference of the prediction accuracy before and after the aforementioned permutation, i.e. $VI^{Tree,j} = pred_{ap}^{(Tree,j)} - pred_{bp}^{(Tree)}$. The variable importance of $\mathbf{x^j}$ across the forest is the mean of the variable importance scores per tree (*Tree*) as $VI^j = \frac{\sum_{Tree=1}^{N_{Tree}} VI^{Tree,j}}{N_{Tree}}$.

This method is specific to Random Forests, as it is based on bisecting the space with the cutpoints extracted during the building process of the forest. Furthermore, it does not provide statistical guarantees.

DeepPINK

Drawing on the insights of *knockoffs*, Lu et al. [2018] exhibited feature selection in deep neural networks by means of *pairwise competition*. Both the original variables $\mathbf{X}$ and their corresponding knockoffs $\tilde{\mathbf{X}}$ are fed to an MLP model augmented using a "pairwise-connected" layers with linear activation, also called *filters*.

Let $\mathbf{Z} \in \mathbb{R}^{p \times 2}$ and $\mathbf{w} \in \mathbb{R}^{p \times 1}$ be the weights connecting the variables and their counterparts to the filters and to the output $\mathbf{y}$ through the MLP hidden layers respectively. The importance measures $\mathbf{z_{imp}}$ and $\tilde{\mathbf{z}}_{\mathbf{imp}}$ are defined as:

$$[\mathbf{z_{imp}}, \tilde{\mathbf{z}}_{\mathbf{imp}}] = \mathbf{Z} \times \mathbf{w}$$

highlighting the competition of each variable against its own knockoff counterpart and the variables against each other. The importance scores are denoted as $\mathbf{VI} = \mathbf{z_{imp}}^2 - \tilde{\mathbf{z}}_{\mathbf{imp}}^2$. This method controls the false discovery rate (FDR) but does not provide a quantitative measure of the importance of individual variables.

2.2.2 . Model Agnostic Methods - Instance based

A popular approach to interpret black-box predictive models is based on *locally interpretable*, i.e. *instance-based*, models. Examples include *SHAP* [Burzykowski, 2020] which is a popular package that measures *local* feature effects using the Shapley values from coalitional game theory. Additional information can be found in the subsequent paragraphs.

Accumulated Local Effects (*ALE*)

Accumulated Local Effects [Apley and Zhu, 2019] is an *instance*-based method that describes how variables influence the prediction of a machine learning model on average. It is mainly used when predictor variables are correlated. *ALE* plots are used to visualize how the predicted outcome changes as a variable's value changes while accounting for interactions with other features. The key idea is to accumulate the average model prediction differences over small intervals of the variable of interest.

The observed values of the variable of interest $\mathbf{x^j}$ are divided into K intervals $N^j(k) = (z_{k-1}^j, z_k^j]$ $(k = 1, \ldots, K)$. The effect of the feature j on the prediction of one instance is computed as:

$$\forall i \in [\![n]\!], \quad \hat{g}_{AL}^j(\mathbf{x_i}) = \sum_{k=1}^{k^j(x)} \frac{1}{n^j(k)} \sum_{i:x_i^j \in N^j(k)} \{\hat{\mu}(\mathbf{x_i^{j|=z_k^j}}) - \hat{\mu}(\mathbf{x_i^{j|=z_{k-1}^j}})\}$$

where $k^j(x)$ is the index of interval $N^j(k)$ in which $\mathbf{x}$ falls, i.e. the j^{th} component of $\mathbf{x_i}$ belongs to this interval, $n^j(k)$ is the number of observations x_i^j belonging to $N^j(k)$ and $\mathbf{x_i^{j|=c}}$ corresponds to the same observation by replacing the j^{th} component with the constant c. The result is a (n, p) matrix denoted $\mathbf{ALE^m}$. Finally, The importance score for the j^{th} component is computed as the average of $\mathbf{ALE^m}$ per variable, $VI^j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{ALE_i^m}$. This method does not provide statistical guarantees.

SHapley Additive exPlanations (*SHAP*)

Shapley [1952] introduced Shapley values as a means derived from cooperative game theory, which aims to fairly allocate the contributions of a group of players in a cooperative game. Štrumbelj and Kononenko [2010] used the Shapley values with the goal of assigning the impact of each input feature in the cooperative prediction per instance of interest. Lundberg and Lee [2017] provided an efficient implementation of this approach with Shapley Additive exPlanations (*SHAP*) [Burzykowski, 2020]. As an *instance*-based method, an aggregation is integrated at the final phase, thereby promoting the method to the *population* level.

First, the dataset is divided following a train/test scheme. Each observation in the test set is considered as an instance of interest. Following this, the *Shapley values* per instance of interest $\mathbf{x}^*$ are computed according to

$$\phi(\mathbf{x}^*, j) = \frac{1}{p}\sum_{s=0}^{p-1} \sum_{\substack{S \subseteq \{1,\ldots,p\} \setminus \{j\} \\ \#S = s}} \binom{p-1}{s}^{-1} \Delta^{j|S}(\mathbf{x}^*)$$

where S is a possible subset of explanatory variables excluding the j^{th} com-

ponent of size s and

$$\Delta^{j|S}(\mathbf{x}^*) = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} \hat{\mu}(\mathbf{x_i}|\mathbf{x_i^{S_1}} = \mathbf{x}^{*,\mathbf{S_1}}, \mathbf{x_i^{S_2}} = \mathbf{x}^{*,\mathbf{S_2}}, \ldots, \mathbf{x_i^{S_{\#S}}} = \mathbf{x}^{*,\mathbf{S_{\#S}}}, x_i^j = x^{*,j}) - \hat{\mu}(\mathbf{x_i}|\mathbf{x_i^{S_1}} = \mathbf{x}^{*,\mathbf{S_1}}, \mathbf{x_i^{S_2}} = \mathbf{x}^{*,\mathbf{S_2}}, \ldots, \mathbf{x_i^{S_{\#S}}} = \mathbf{x}^{*,\mathbf{S_{\#S}}}).$$

SHAP measures the conditional effect of including a particular variable on the model's output, while considering interactions with other variables. The result is a (n_{test}, p) matrix. Finally, the average of the *Shapley values* per variable is computed. This method does not offer statistical guarantees.

2.2.3 . Model Agnostic Methods - Population level

Global, i.e. population-level, explanations are better suited than instance-level explanations for epidemiological studies and scientific discovery in general.

Mean Decrease Accuracy (*MDA*)

First introduced by Breiman [2001], *MDA* grabbed the attention of multiple researches for its simple principle and efficient implementation [Bracher-Smith et al., 2022, Covert et al., 2022, Debeer and Strobl, 2020, Nicodemus et al., 2010].

To retrieve the impact of one variable on the prediction of the outcome, the key idea was first, to break the relation between this variable and the outcome by integrating some perturbation within the variable (e.g. permutation) followed by the comparison of the loss function before and after this applied perturbation. A severe drop in the performance reflects the degree of importance of the variable of interest for the prediction.

Another recent paper by Mi et al. [2021] has introduced model-agnostic explanation for black-box models based on the *permutation* approach. *Permutation* importance can work with any learner. Moreover, it relies on a single model fit, hence it is an efficient procedure. In order to ascertain the significance of the variable of interest x^j, a series of permutations are conducted, with the importance score computed according to $\hat{m}^j = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left((y_i - \hat{\mu}(\mathbf{x_i^{(j)}}))^2 - (y_i - \hat{\mu}(\mathbf{x_i}))^2 \right)$. Subsequently, the Wald statistic $z^j = \frac{mean(\hat{m}^j)}{std(\hat{m}^j)}$ is derived under the assumption that it follows a standard normal distribution. Finally, the corresponding p-values can be calculated. This method offers statistical guarantees by controlling the type-I error rate for the significance of individual variables in settings where these variables are not correlated.

Shapley Additive Global importancE (*SAGE*)

While *SHAP* focuses on the *local interpretation* by aiming to explain a model's individual predictions and the need for a post-aggregation step to promote the method to the *population* level, *SAGE* [Covert et al., 2020, Kumar et al., 2020] is an extension to *SHAP* assessing the role of each feature in a *global interpretability* manner. The SAGE values are derived by applying the Shapley value to a function that represents the predictive power contained in subsets of features as

$$\phi(j) = \frac{1}{p} \sum_{S \subseteq \{1,\dots,p\} \setminus \{j\}} \binom{p-1}{\#S}^{-1} MI(\mathbf{y}, \mathbf{X}^{S \cup \{j\}})$$

where *MI* represents the mutual information i.e. the decrease in $\mathbf{y}$ uncertainty when integrating the j^{th} component into different subsets S. This method does not provide statistical guarantees.

Knockoff

As described in [Candes et al., 2017, Barber and Candès, 2015], the knockoff filter is a variable selection method for multivariate models that controls the False Discovery Rate (FDR).

The knockoffs $\tilde{\mathbf{X}}$ are a family of random variables with two important properties:

$$\forall S \subset [\![p]\!], (\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap(S)}} \overset{d}{=} (\mathbf{X}, \tilde{\mathbf{X}}) \quad \text{and} \quad \tilde{\mathbf{X}} \perp\!\!\!\perp \mathbf{y} | \mathbf{X}.$$

Thus, the first step of this procedure involves sampling extra null variables that have a correlation structure similar to that of the original variables. A statistic is then calculated to measure the strength of the original variables versus their knockoff counterpart with *l1-regularization path* or *cv lasso*.

For *l1-regularization path*, the following problem is solved:

$$\underset{\mathbf{w}}{\text{argmin}} \frac{1}{2n} ||\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\mathbf{w}||_2^2 + \lambda ||\mathbf{w}||_1 \tag{2.5}$$

through a set of λ entries. If the corresponding coefficient $w^j > 0$, the maximum λ value is returned.

For *cv lasso*, the same problem in (2.5) is solved without the regularization component where a vector $\mathbf{w}$ of coefficients is returned. Subsequently, the importance scores are set to as $\mathbf{VI} = |\mathbf{w}_{:p}| - |\mathbf{w}_{p:}|$ where p is the number of features.

Finally, a recent paper by [Watson and Wright, 2021] showed the necessity of conditional schemes and introduced a knockoff sampling scheme, whereby the variable of interest is replaced by its knockoff to monitor any drop in performance of the leaner used without refitting. This method is computationally inexpensive, and enjoys statistical guarantees from [Lei et al., 2018]. However

it depends on the quality of the knockoff sampling where even a relatively small distribution shift in knockoff generation can lead to large errors at inference time. Although knockoffs are subject to randomness due to sampling, they permit control over the false discovery rate (FDR) of the selection procedure. However, they do not provide a quantitative measure of individual variable importance.

Leave One Covariate Out (*LOCO*)

Rather than performing a perturbation to the variable of interest in order to break its relation with the outcome for a model fitted once, the complete removal of the variable was proposed by [Williamson et al., 2021, Lei et al., 2018], while *refitting* the core learner.

The importance of variable $\mathbf{x^j}$ is measured as the change in the performance by comparing the full model with the refitted model excluding the variable of interest as

$$VI^j = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} L(\hat{\mu}(\mathbf{x_i}), y_i) - L(\hat{\mu}(\mathbf{x_i^{-j}}), y_i)$$

where L the loss function. The p-values are computed under the standard Gaussian distribution assumption. This method does provide statistical guarantees in terms of controlling the type-I error rate for the significance of individual variables.

Regardless of whether they show the asymptotic consistency of the model, their approach is intractable, given that it requires refitting the model for each variable.

Gao et al. [2022] applied the aforementioned removal method to a DNN model by integrating linear assumptions to accelerate the training process. However, statistical guarantees may be lost due to the alteration of variable importance in the linearized model.

2.3 . Conclusion

A multitude of methodologies have been proposed for the assessment of the relative importance of individual variables across diverse contexts. Conversely, other studies have concentrated on the comparison of specific models within distinct communities [Altenmüller et al., 2021, Liu et al., 2021, Mi et al., 2021, Janitza et al., 2018, Chipman et al., 2010]. However, these comparisons lack conceptualization from a unified perspective.

Furthermore, previous work has established potential advantages of conditional permutation schemes for inference of variable importance. Although there is a clear need to understand how different permutation schemes impact model performance, this has not been extensively explored due to a lack

of efficient computational methods. This limitation impedes researchers from conducting a comprehensive comparison of these schemes with other techniques across a broader spectrum of predictive modeling scenarios.

3 - Group Variable Importance Methods

Summary Different methods have been presented in the previous chapter to assess the importance of individual variables across different domains and understand their impact on the decision-making process. In high-correlation settings, conditional schemes such as conditional permutation were proposed as a solution to the appearance of *fake relevant* variables highly correlated with the true relevant ones. Nevertheless, given two extremely-correlated variables, the conditional permutation results in the mutual cancellation of the conditional importance scores due to a limitation in the definition. Therefore, moving from the single to the group level have been proposed as a remedy. Additionally, when dealing with high-dimensional settings having hundreds or thousands of variables (example in Neuroimaging), grouped interpretations are preferred to reduce the need for expensive computations. Moreover, when dealing with categorical variables that have been dummy or one-hot encoded, it is important to treat them as a single group in order to maintain relationship between the different categories. Before applying group-based methods, it is necessary to *group variables* based on some domain knowledge in a particular field or shared statistical properties derived from the data. One initial thought is to consolidate each group into one individual variable via *summarization* or *stacking*. To gain deeper insights at the group-level, various perspectives on importance scoring methods are available. On one hand, *grouped non-perturbative methods* rely on the architectual or statistical properties of the method such as *Marginal Importance* or *Random Forests*. On the other hand, *grouped removal-based methods* involve perturbations, such as complete removal, exposed to the group of interest within *one fit* or a *refitting* process respectively. In this chapter, we study the literature of the state-of-the-art methods for group-based variable importance. Table 3.1 provides a summary of the methods employed in this chapter. These include model-agnostic vs model-specific approaches, removal-based vs non-perturbative approaches, global-based vs instance-based approaches, which provide or do not provide statistical guarantees.

3.1 . Variables Grouping

Before deploying group-level methods, the first step is to *group variables*. This can be done using a custom *knowledge-driven* approach based on specialized knowledge within a particular field. Conversely, the absence of this information leads to a *data-driven* approach where the degree of correlation serves as the splitting criterion. In both cases, an instance may belong to one or more groups depending on the technique used.

3.1.1 . Data-driven grouping (Clustering Methods)

Agglomerative Clustering [Müllner, 2011, Johnson, 1967]

By considering each instance as its own cluster, the method is based on an iterative merging of the closest pairs of clusters to compose larger ones. The merging process continues until reaching a predetermined number of clusters or until a certain linkage criterion (e.g., distance threshold) is satisfied.

Divisive Clustering [Reddy et al., 2017, Kaufman and Rousseeuw, 1990]

Instead of merging pairs of clusters, this method considers all the observations attached to one big cluster followed by an iterative division into smaller clusters until the pre-defined condition is reached. It constructs a hierarchy of clusters in a dendrogram, with each level representing a different partitioning of the data

Fuzzy Clustering [Jaeger et al., 2003]

Rather than taking part in one cluster as in the aforementioned techniques, *Fuzzy clustering* allows instances to belong to multiple clusters under varying degrees when these instances may exhibit ambiguity or uncertainty regarding their cluster assignments.

However, these groups depend only on the statistical similarity which might not coincide with domain-specific interpretations [Chakraborty and Pal, 2008].

K-means Clustering [MacQueen, 1967]

K-means clustering partitions a dataset into a pre-specified number (k) of clusters by minimizing the within-cluster variance. Each data point is assigned to the cluster with the closest centroid, which represents the mean of the points within that cluster. The algorithm iteratively refines the cluster centroids and reassigns data points until a convergence criterion is met, typically minimal changes in cluster assignments or centroid locations. While computationally efficient and straightforward to implement, k-means clustering exhibits sensitivity to the initial selection of cluster centroids and can converge to locally optimal solutions that may not be globally optimal.

3.1.2 . Knowledge-driven grouping

Variables are grouped based on their thematic attributes (e.g., measurement device) rather than on patterns or statistical properties observed in the data. This grouping can enhance the effectiveness and interpretability of the grouping process, leading to more meaningful insights. An illustration is the *PhoneStudy* dataset in the work by Stachl et al. [2020] where the variables have been categorized based on specific behaviors, such as app usage, music consumption, or overall phone usage.

3.2 . Solo Group Representative

Given that the variables are associated with a cluster, it is possible to identify a single representative of each group via *aggregation* or *stacking*. Consequently, by returning to the single level, the application of the methods elaborated in the previous section is now feasible.

3.2.1 . Cluster Summarization via Aggregation

[Aggarwal and Han, 2014] One intuitive idea is to apply an aggregation step over the data points in a cluster by computing the mean, median or maximum values across the variables of the group of interest.

3.2.2 . Stacking Approach

A different angle can be motivated by a recent line of work that developed model-stacking techniques [Wolpert, 1992] which combine different input domains and groups of variables rather than aggregating different estimators on the input data. This approach has been used in various applications ranging from video analysis [Zhou et al., 2021] over protein-protein interactions [Albu et al., 2023] to neuroscience applications [Rahim et al., 2015].

A key benefit of multimodal or group stacking is that it allows for modality-specific encoding strategies and while approaching inference at the simplified level of the 2^{nd} level model combining the modality-wise predictions or activations. This strategy has been used to explore importance of distinct types of brain activity at different frequencies for age prediction [Sabbagh et al., 2023, Engemann et al., 2020].

Method	Model-agnostic?	Removal based?	Global?	Statistical guarantees?
Grouped Marginal	No	No	Yes	FPR
Grouped MDI	No	No	Yes	No
Grouped Shap	Yes	Yes	No	No
GPFI	Yes	Yes	Yes	No
GOPFI	Yes	Yes	Yes	No
Grouped Perturbations	Yes	Yes	Yes	FDR
LOGO	Yes	Yes	Yes	No
LOGI	Yes	Yes	Yes	No

Table 3.1: **Summarizing table for group-based methods**: This table provides a summary of the methods presented in this chapter, categorizing them into three main groups: model-agnostic vs model-specific, removal-based vs non-perturbative and global vs instance-based approaches. It indicates whether each method provides statistical guarantees or not. *FDR*: False Discovery Rate. *FPR*: False Positive Rate.

3.3 . Grouped Non-perturbative Methods

Grouped Marginal Importance

Given the necessity for a reference point when evaluating various methods, we incorporate the *Grouped Marginal Importance* scores of the various predictors. Each group $\mathbf{X}^{\mathcal{J}} \in \mathbb{R}^{n \times \#\mathcal{J}}$ is fitted separately to predict the response $\mathbf{y}$ using a *Multivariate Linear Regression* model as:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{X}^{\mathcal{J}}$$

The *F-test* ($\mathcal{H}_0 : \beta_1 = 0$ vs $\mathcal{H}_1 : \beta_1 \neq 0$) was used to extract the corresponding p-values. The F-statistic is computed as $F = \frac{\text{Mean Squared Regression}(MSR)}{\text{Mean Squared Error}(MSE)}$ where $MSR = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\#\mathcal{J}}$ having $\bar{y}$ the mean of the observed values of the outcome y and $MSE = \frac{\sum_i^n (y_i - \hat{y}_i)^2}{n - \#\mathcal{J} - 1}$. Using the computed F-score and the degrees of freedom of the numerator ($\#\mathcal{J}$) and the denominator ($n - \#\mathcal{J} - 1$) from the *F-Table*, we can extract the corresponding p-value denoted as $p^{\mathcal{J}}$. This method offers statistical guarantees by controlling the type-I error rate for the significance of groups.

Grouped Mean Decrease Impurity

Wehenkel et al. [2018] have introduced group-based variable importance for Random Forests, extending the seminal work of Louppe et al. [2013] on *Mean Decrease Impurity (MDI)* presented in section (2.1.1). A feature is defined as important by appearing more often in the forest and at top nodes and/or by strongly reducing the impurity at the nodes where it belongs.

Once all the variables have their corresponding impurity function scores, the importance score of the group of interest follows three aggregation functions of the impurity scores among the participating variables:

- The sum where the significance of the group of interest is determined by the total reduction in class impurity achieved by the variables it contains.
- The average which prevents bias arising from variations in group size.
- The maximum where the variable deemed most important serves as the sole representative of the group's significance.

Despite that, (1) the sum displays bias in favor of larger-sized groups, (2) the average diminishes a group's significance when only a small fraction of its variables hold importance and (3) the maximum suggests that one variable reflects the collective importance of the group. Also, this method does not offer statistical guarantees.

3.4 . Grouped Removal-based Methods

3.4.1 . Grouped Shapley Values

Built upon *SHAP* detailed in section (2.2.2) that has been introduced to explain the prediction of a new instance on the single-variables level, Jullum et al. [2021] altered the definition of the offered explanation so as to shift to the group level.

In this work, the shapley value of the group of interest $\mathcal{J}$ is defined as

$$\phi(\mathbf{x}^*, \mathcal{J}) = \sum_{\tau \subseteq \mathcal{S} \setminus \mathcal{J}} \frac{\#\tau!(\#\mathcal{G} - \#\tau - 1)!}{\#\mathcal{G}!}(\upsilon(\tau \cup \mathcal{J}) - \upsilon(\tau))$$

$$\upsilon(\tau) = \mathbb{E}[\hat{\mu}(\mathbf{x})|\mathbf{x}^{\tau} = \mathbf{x}^{*\tau}]$$

where τ runs over the groups (not the individual variables) in $\mathcal{S}$.

An alternative approach is to simply aggregate the single Shapley values in order to obtain the related explanations of the groups, i.e. $\phi^{\mathcal{J}} = \sum_{j \in \mathcal{J}} \phi^j$. This approach is equivalent to the previous approach in certain circumstances.

Serving as an instance-based method, an aggregation step is applied to retrieve the population-level explanations of the different groups or *SAGE* detailed in section (2.2.3) can be utilized once its definition has been upgraded. this method does not offer statistical guarantees.

3.4.2 . Grouped Permutation-based Methods

Grouped Permutation Feature Importance (GPFI)

Mi et al. [2021] proposed an efficient model-agnostic procedure for black-box models' interpretation. It uses the *permutation approach* [Fisher et al., 2019, Breiman, 2001] with the importance score computed as the reduction in a model's performance when randomly shuffling the variable of interest. To account for group-level structure, Gregorutti et al. [2015] suggested taking into account all the variables of the group of interest in the permutation scheme jointly, known as *Group Permutation Feature Importance(GPFI)*.

$$VI^{\mathcal{J}} = \mathbb{E}(L(\hat{\mu}(\tilde{\mathbf{x}}^{\mathcal{J}}, \mathbf{x}^{-\mathcal{J}}), y)) - \mathbb{E}(L(\hat{\mu}(\mathbf{x}), y))$$

Grouped Only Permutation Feature Importance (GOPFI)

Au et al. [2021] proposed *Group Only Permutation Feature Importance (GOPFI)* which examines the level of the group's individual contribution to the model's performance. The random joint shuffling is performed for all the variables of the different groups expect the ones of the group of interest.

Hence, a group of interest $\mathcal{J}$ is regarded as relevant by observing a decrease in the expected loss between the joint permutation of all the variables and the joint permutation except the considered group cases.

$$VI^{\mathcal{J}} = \mathbb{E}(L(\hat{\mu}(\tilde{\mathbf{x}}), y)) - \mathbb{E}(L(\hat{\mu}(\mathbf{x}^{\mathcal{J}}, \tilde{\mathbf{x}}^{-\mathcal{J}}), y))$$

However, according to Strobl et al. [2008], simple permutation approaches yield poor accuracy and specificity in high correlation settings.

Grouped Global Perturbations

Whereas the aforementioned methods apply the joint permutation of all the variables of the group of interest, Lee et al. [2018] proposed a broader concept for the applied perturbations to the variables and groups of interest while providing p-values under hypothesis testing.

Nevertheless, they did not address the degree of correlation between the variables (and the groups), which increases the difficulty of the problem. This method offers statistical guarantees by controlling the false positive rate (FDR) while providing a quantitative measure of the importance of individual groups in terms of p-values.

3.4.3 . Grouped Refitting-based Methods

Leave One Group Out (LOGO)

While Williamson et al. [2021] proposed a model-agnostic approach based on refitting the learner after the removal of a variable of interest also called *LOCO (Leave-One-Covariate-Out)* by Lei et al. [2018], Au et al. [2021] have adapted this work to the group-level by considering the removal of all the variables of the group of interest jointly.

Thus, a group of interest $\mathcal{J}$ is considered as relevant when observing an increase in the expected loss compared to the full model's expected loss when leaving out a group of variables and performing a refit. The importance score is denoted as:

$$VI^{\mathcal{J}} = \mathbb{E}(L(\hat{\mu}(\mathbf{x}^{-\mathcal{J}}), y)) - \mathbb{E}(L(\hat{\mu}(\mathbf{x}), y))$$

This method does not offer statistical guarantees.

Leave One Group In (LOGI)

In lieu of removing the group of interest $\mathcal{J}$ as mentioned in section (3.4.3), Au et al. [2021] established *Leave-One-Group-In (LOGI)* that assesses the solo impact of the group of interest on the prediction when removing all the remaining variables.

As a result, a group of variables (e.g. all measurements from a specific medical device) is deemed as relevant when reducing the expected loss in contrast to the null model. The importance score is defined as:

$$VI^{\mathcal{J}} = \mathbb{E}(L(\hat{\mu}_{null}, y)) - \mathbb{E}(L(\hat{\mu}(\mathbf{x}^{\mathcal{J}}), y))$$

where $\hat{\mu}_{null}$ provides the prediction as the average of the outcome.

However, both approaches become intractable due to the necessity of refitting the learner for each group, particularly in the case of low cardinality groups. This method does not offer statistical guarantees.

3.5 . Conclusion

The utilization of various methods to perform group-based variable importance analysis via simple fit perturbation or refitting removal has been demonstrated to lack statistical control or to be computationally expensive in high-dimensional cases due to the need of retraining the involved model. Additionally, the identification of conditional variants on the group-level that are both theoretically and empirically demonstrated to be effective has not been identified in the literature. Consequently, there is a necessity for the development of a sensitive, non-linear, agnostic method with statistical guarantees, which is equipped with good computational performance.

While stacking is straightforward to implement with standard software, for instance, scikit-learn [Pedregosa et al., 2011], the inference process with stacking has not yet been formalized. Moreover, it necessitates the fitting of multiple disconnected estimators, which may restrict the capacity of the model.

4 - Variable Importance for Population Imaging in Neuroimaging

Summary The field of interpretable AI/explainable AI (XAI) is a rapidly developing area that bridges the gap between AI and human-computer interaction [Ali et al., 2023]. Its core objective is to make AI systems understandable by humans, thereby fostering trust and transparent decision-making. In the past, researchers have traditionally analyzed brain activity in a marginal manner, examining each voxel individually (mass univariate statistics). This approach has been linked to the marginal importance discussed earlier, with a focus on individual variables. However, this method has limitations in capturing interactions between variables. To address this, clustering statistics or anatomical atlases have been employed to provide regions of interest (ROIs), which do not offer a comprehensive global view of the brain. The preceding chapters provided an overview of the current state-of-the-art methods for variable importance. However, these methods lacked a statistically rigorous groundwork that considers dependencies between variables. Additionally, the chapters addressed the utility of potential alternatives based on conditional inference. Consequently, the current XAI literature may not be sufficiently effective for addressing the challenges encountered with neuroscience or neuroimaging data characterized by both *high-dimensionality* and *high-correlation* among variables. Furthermore, the utility of *brain imaging* in characterizing brain connectivity and structure for the prediction of individual-based traits has been a subject of debate in the literature. Brain imaging modalities have been analyzed to provide additional insights into the complex challenge of comprehending brain and mental disorders. In this chapter, the focus is put on providing the necessary background on the challenges of dealing with population-level datasets and clarifying the necessity of incorporating statistical inference into established ML applications in neuroscience.

4.1 . Brain Imaging Modalities (Neuroimaging)

Neuroimaging encompasses a wide range of brain mapping techniques that are a fundamental tool in the field of cognitive neuroscience to visualize both the structure and function of the brain [Friston, 2009]. These techniques are central to clinical neuroscience for their ability to provide insight into brain health and pathology. For example, Magnetic Resonance Imaging (MRI) has proven its ability to study the details of brain structures in individuals afflicted with neurological or psychiatric disorders [Kolbeinsson et al., 2020]. Clinicians and researchers can gain valuable information for diagnosis, or to unravel the

underlying mechanisms of various diseases. These indicators have proven beneficial in understanding the extent of neurological conditions across various demographic groups, thereby increasing the accuracy of diagnosis and prognosis. Fig 4.1 illustrates the various brain imaging modalities acquired using the brain that will be discussed in greater detail in the following sections.

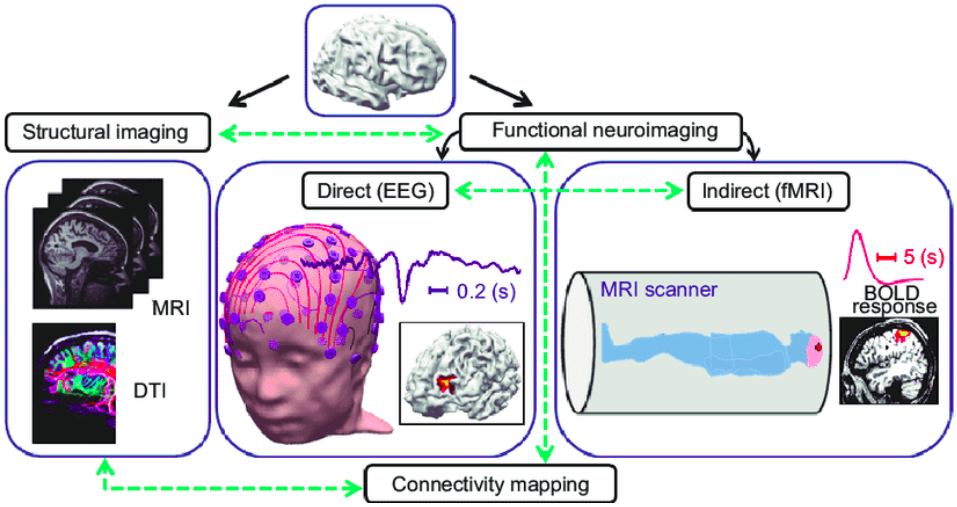

Figure 4.1: **Segmentation of the brain**: This figure presents a variety of neuroimaging modalities that interact with one another to clarfiy underlying brain networks. Figure is derived from [Edelman et al., 2015].

4.1.1 . Magnetic Resonance Imaging (MRI)

Magnetic resonance imaging (MRI) is a non-invasive imaging technique that provides detailed images of the brain's anatomy and function [Edelman Robert R and Warach Steven, 1993, Lauterbur, 1989]. It delineates various types of tissue, such as white and grey matter. Unlike traditional X-rays, MRI does not use ionizing radiation, rendering it a safer option for scrutinizing brain tissue. The MRI machine generates a strong magnetic field, which causes the hydrogen atoms in the body to align in a specific direction. Radio waves are then emitted, causing these aligned atoms to produce signals that are detected by the MRI machine's receiver. By analyzing these signals, the machine constructs highly detailed images of the brain's structures and any abnormalities present. The utilization of different MRI sequences, such as T1-weighted, T2-weighted, and diffusion-weighted imaging, leads to the acquisition of complementary information regarding the brain's tissues and pathology. The acquisition process typically spans between 15 minutes and an hour, depending on the complexity of the scan and the specific protocols employed. MRI is crucial in diagnosing numerous neurological disorders such as tumors, infections, strokes, developmental anomalies, and degenerative

illnesses. Nonetheless, access to MRI technology and participation in neuroscientific research vary across different regions and demographic groups, potentially resulting in selection bias [Fry et al., 2017].

4.1.2 . Functional Magnetic Resonance Imaging (fMRI)

Functional magnetic resonance imaging (fMRI) is a non-invasive neuroimaging technique that is utilized to decipher brain activity patterns by capturing alterations in blood flow and oxygenation levels [Ogawa et al., 1990]. It operates within a magnetic resonance imaging (MRI) scanner, where it employs powerful magnetic fields and radiofrequency pulses to manipulate and detect changes in the alignment of hydrogen atoms in the body, particularly in the brain. When neural activity increases in specific brain regions, it triggers a surge in blood flow to meet the heightened metabolic demands, a phenomenon known as the hemodynamic response. This increase in blood flow results in a corresponding alteration in the magnetic properties of blood, which fMRI equipment can discern with remarkable precision. By analyzing the blood flow signal within each voxel over time, fMRI generates detailed spatial maps of brain activity, unveiling which regions are actively engaged during particular tasks, sensory inputs, or cognitive processes. However, movement within the scanner can disrupt the signal, and the high cost and limited accessibility restrict its use. Consequently, the acquisition process necessitates a series of preprocessing steps, including motion correction, spatial normalization, and temporal filtering. These steps are employed to enhance data quality and extract reliable signals from the raw data. Over time, fMRI has become an indispensable tool in neuroscience research, enabling scientists to explore the intricacies of brain function, map neural networks, investigate neurological disorders, and evaluate the effects of interventions or treatments on brain activity and connectivity.

4.1.3 . Electrophysiological Methods (M/EEG)

Magnetoencephalography (MEG) [Cohen, 1972] and electroencephalography (EEG) [Berger, 1969] are non-invasive neuroimaging techniques used to offer insights into brain health by directly measuring the electrical activity of the brain without penetrating the skull. EEG records the electrical signals produced by neurons using electrodes placed on the scalp, while MEG measures the magnetic fields generated by these electrical currents using sensors placed around the head. Both MEG and EEG provide millisecond temporal resolution, allowing researchers to capture the rapid dynamics of neural activity associated with various cognitive processes, sensory inputs, and motor functions. The acquisition of MEG and EEG involves positioning the sensors or electrodes on the scalp according to a specific layout, typically based on international standards such as the 10-20 or 10-5 systems for EEG and a helmet-shaped array for MEG. Signal preprocessing steps, including artifact removal,

noise reduction, and spatial filtering, are applied to enhance the quality of the data. The combined use of MEG and EEG offers unique benefits. MEG provides excellent spatial resolution, although it is sensitive to superficial sources [Hämäläinen et al., 1993]. EEG, on the other hand, is more sensitive to deeper brain activity but has poorer spatial resolution [Luck, 2005]. Collectively, these techniques offer valuable insights into the dynamics of brain function and are widely used in cognitive neuroscience, clinical research, and brain-computer interface applications. MEG and EEG enable the assessment of brain health on a broad scale, potentially advancing preventive public health initiatives.

4.1.4 . Combining brain imaging modalities

Electroencephalography (EEG) has demonstrated its utility across a broad spectrum of specialized domains, encompassing surgical environments [Samanta, 2022] and sleep studies [Desjardins et al., 2017]. Although both EEG and MEG serve as valuable instruments, their capacity to capture intricate anatomical details is limited. Consequently, clinical research in neurology frequently combines EEG and MEG with complementary neuroimaging modalities possessing superior spatial resolution. Such modalities include magnetic resonance imaging (MRI), and functional magnetic resonance imaging (fMRI). The integration of expert-derived features derived from all three modalities (M/EEG, fMRI, and MRI) can enhance the learning algorithms utilized to identify surrogate biomarkers, potentially leading to more robust and informative brain assessments.

4.2 . Extending the reach of neuroscientific research with interpretable machine learning

Neuroscience plays a pivotal role in elucidating the workings of the human mind and nervous system. It employs a diverse array of brain modalities, including high-resolution imaging techniques such as MRI and electrophysiological recordings such as M/EEG, to investigate the intricate structure and function of the brain. Its investigations have the potential to revolutionize our understanding of consciousness, behavior, and cognition. This, in turn, could pave the way for advancements in healthcare and our overall grasp of the human experience and the societal burden of brain disorders such as Alzheimer's dementia[1] or epilepsy[2]. Neuroscience has historically employed simple, transparent models for the analysis of brain data. These models, such as linear regressions or ANOVA, are relatively straightforward to use thanks to modern software; these models have quite some heavy theory behind them,

[1] https://www.who.int/news-room/fact-sheets/detail/dementia

[2] https://www.who.int/news/item/20-07-2023-new-global-action-plan-on-epilepsy-and-other-neurological-disorders-published

and it is possible to ascertain the direct contribution of each variable in the model (e.g., the activity of a specific brain region) to the final result (e.g., the prediction of memory performance). In light of the previous utilization of smaller datasets and the reliance on clustering statistics to group interconnected variables, the marginal importance approach continues to be a valuable tool. However, the complex nature of the brain, with its complex neural networks comprising numerous connections between neurons and dynamic processes [Bassett and Gazzaniga, 2011], and its integration of information from various sources, which results in highly correlated inputs [Eggermont, 1990], presents challenges for simple models [Badrulhisham et al., 2024]. Using such models, it is difficult to capture the complex interactions and hidden patterns [Räz, 2024]. To address this issue, researchers have employed grouping techniques and brain maps to identify specific regions of interest (ROIs). The selection of regions of interest (ROIs) presents a challenging decision, as the optimal selection may differ depending on various conditions or pathologies [Smith et al., 2011]. Pre-defined reference anatomical atlases, such as Automated Anatomical Labeling (AAL) [Tzourio-Mazoyer et al., 2002] or sulci-based atlases [Perrot et al., 2009, Desikan et al., 2006], provide examples of such resources. However, these methods do not provide a comprehensive understanding of the entire brain. As illustrated in Fig. 4.3, there are around

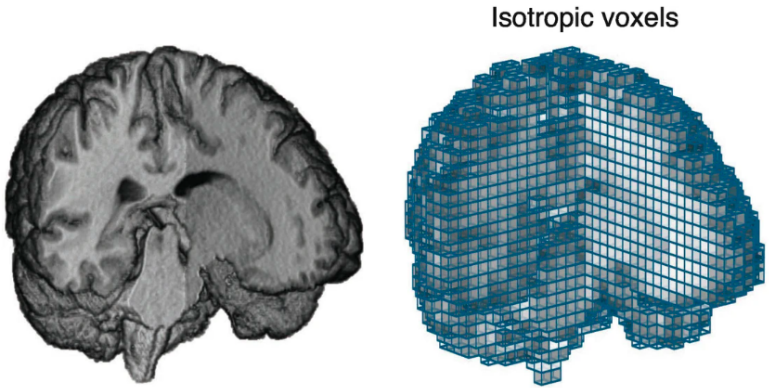

Figure 4.2: **Segmentation of the brain**: This figure presents a decomposition of the brain into voxels. Left: volume rendering of an excavated T1-weighted MR image. Right: voxel grid with isotropic, i.e., cubic, voxels overlaid on the MRI. Figure is derived from [Burgos, 2023].

120 billions of neurons in the human brain [Herculano-Houzel, 2009], yet only hundreds of thousands of voxels are analyzed. This demonstrates the necessity for multivariate models combined with multimodal data to gain more profound insights with larger datasets, which are necessary to capture hidden brain functions [Woo et al., 2017]. In addition, Fig. 4.2 highlights the challenge of brain imaging data through parcellations and correlation connectivity

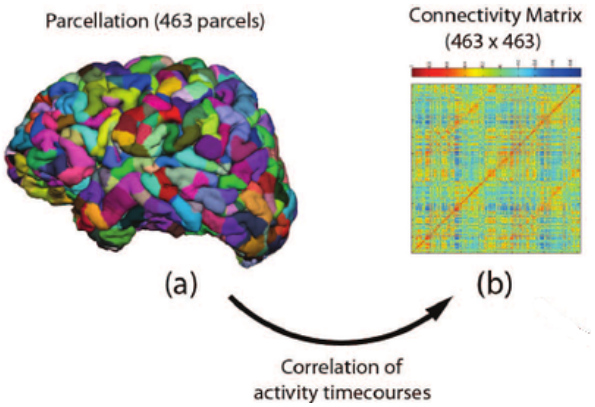

Figure 4.3: **Whole-brain functional connectivity data**: This figure visualizes the data derived from fMRI timecourses in cortical areas (a), and calculation of a correlation matrix between every pair of these regions (b). Figure is derived from [Böttger et al., 2014], © [2014] IEEE.

matrices, which requires the application of processing pipelines compared to tabular data.

Machine Learning (ML) is a promising tool for the discovery of novel biomarkers from heterogeneous data sources, including omics, imaging and multimodal data. This could lead to the development of new medications and personalized diagnostics and treatments [Dara et al., 2022, Newby et al., 2021]. This is particularly important in the context of brain health and the global effort to develop medicines for brain disorders, including dementias [Shan and Lagopoulos, 2023, Singh et al., 2022]. The expansion of biobanks and the advancement of statistical learning techniques have led neuroscientists to increasingly employ sophisticated machine learning algorithms to analyze brain activity. However, the complex machine learning models employed to perform this analysis frequently exhibit a notable absence of interpretability, which remains paramount in neuroscience [Hofmann et al., 2022]. A model that simply produces outcomes without clarifying the underlying reasoning, commonly referred to as a "black box," obstructs scientific understanding and limits its usefulness in clinical contexts [McKelvey et al., 2018]. XAI facilitates the unpacking of these "black box" models, thereby revealing the underlying reasoning that leads to conclusions about brain function or disease risk. This high transparency is of paramount importance in medicine, where decisions based on AI can have significant consequences. By giving importance to both interpretability and the power of machine learning, neuroscientists can study the brain while upholding scientific integrity and instilling trust in their discoveries.

With regard to the challenge of predicting at the individual level using diverse neuroimaging modalities, each governed by unique data-generating

mechanisms, recent advancements have employed model-stacking techniques [Karrer et al., 2019, Liem et al., 2017]. Rahim et al. [2015] studied classification of Alzheimer's diagnosis by merging fMRI and PET data through a stacking approach [Wolpert, 1992], where the stacked models utilized input from various modalities. Subsequently, Liem et al. [2017] employed a comparable approach for age prediction and observed that integrating anatomical MRI with fMRI markedly reduced errors while enhancing the identification of cognitive impairment.

In high-dimensional neuroscience settings, where there is a vast array of brain phenotypes, the process of iterating through all variables is both time-consuming and almost impractical. Consequently, the necessity for grouping in high-dimensional settings becomes apparent. Finally, there is a need for a robust statistical model with high capacity, at both single and group levels, that considers the interconnections between variables with a preference for *agnostic* interpretability. This is because researchers want to have insights into their models based on tested principles.

4.3 . Proxy Measures

Proxy measures are indirect indicators or substitutes employed to estimate or infer the value of an underlying construct or phenomenon that cannot be directly measured [Roydhouse et al., 2022, Hrisos et al., 2009]. These measures are often employed when direct assessment is impractical, expensive, or impossible. Proxy measures rely on the assumption that they are correlated with the target variable of interest, allowing researchers to make inferences about the target variable based on the observed proxy. For instance, income level may be employed as a proxy measure for socioeconomic status, while heart rate variability could serve as a proxy for stress levels. Proxy measures are frequently utilized in various fields, including economics, sociology, and public health, to study complex phenomena and inform decision-making in the absence of direct measurements.

In recent years, there has been a notable increase in interest surrounding the concept of brain age as a potential indicator of brain health [Sone and Beheshti, 2022]. This interest has prompted the development of the brain age delta, which highlights the difference between an individual's predicted brain age and their actual chronological age [Smith et al., 2019]. Studies have indicated that this brain age delta may serve as an indicator of both physical and cognitive decline in adults, offering insights into neurodegenerative processes [Liem et al., 2017]. The encouraging findings with brain age as a brain-derived biomarker necessitated further investigation beyond the construct of pathological aging.

Dadi et al. [2021] predict the chronological age with fMRI modality and

socio-demographic data as inputs. The authors demonstrate that the application of machine learning to population modeling enables the extraction of mental health indicators from a range of sources, including brain signals and questionnaire responses. These derived measures have the potential to enhance or even replace traditional psychometric evaluations within clinical populations. Nevertheless, the researchers lacked the requisite tools to assess the true impact of each brain modality or sociodemographics on the prediction of brain age. This is due to the current limitations status of XAI, which lacks a high-capacity, statistically rigorous method that can effectively control for the appearance of spurious relevant predictors. Furthermore, it is not always clear that brain data represents the most crucial source of information, particularly when complex psychological traits or social outcomes are predicted. Variable importance can help separating signals and clarifying what is not redundant.

4.4 . Conclusion

While machine learning offers a suite of sophisticated algorithms for gaining deeper insights into the complex nature of the brain through the use of brain imaging modalities and socio-demographic data, these models are still "black boxes" from a human perspective. Interpretable AI/explainable AI (XAI) has developed a range of statistical-based explanation methods to enhance the transparency of such models, thereby building trust and accountability for the decision-making processes. Nevertheless, this field lacks a sensitive, non-linear, agnostic method with statistical guarantees to provide statistically valid insights across different applications. The absence of suitable tools has resulted in the inability to ascertain the true impact of brain variables.

4.5 . Scientific goals of the thesis

The discussion up to now shows that the complexity of the human brain together with the high-dimensionality of neuroscience methods poses special challenges to ML/XAI methods, which I wish to study in this thesis. In light of these challenges, the following contributions have been made:

1. Find algorithm that is expressive to detect important variables in non-linear models.
2. Deal with extremely correlated variables in high-dimensional settings.
3. Revisit previous applied ML literature to investigate the potential impact of statistically controlled methods.

Part II

Contributions

Preliminaries

Experiments Setting

In the following chapters, in order to ensure a fair comparison across experiments, all methods are employed with their original implementations. As for {*CPI*, *BCPI*}*-DNN*, {*CPI*, *BCPI*}*-RF* and {*Permfit*, *BPI*}*-DNN* particularly, the default behavior consists of a 2-fold internal cross validation where the importance inference is performed on an unseen test set. The scores from the various splits are then aggregated to compute the final variable importance. All experiments are conducted with 100 runs.

Evaluation Metrics

AUC score [Bradley, 1997] The variables are ordered by increasing p-values, yielding a family of p splits into relevant and non-relevant at various thresholds. AUC score measures the consistency of this ranking with the ground truth ($p_{signals}$ predictive features versus $p - p_{signals}$).

Type-I error Some methods output p-values for each of the variables, that measure the evidence against each variable being a null variable. This score checks whether the rate of low p-values of null variables exceeds the nominal false positive rate (set to 0.05).

Power This score reports the average proportion of informative variables detected (when considering variables with p-value < 0.05).

Computation time The average computation time per core on 100 cores.

Prediction Scores As some methods share the same core to perform inference and with the data divided into a train/test scheme, we evaluate the predictive power for the different cores on the test set.

5 - Statistical Valid Importance: the Case of Single Variables

Summary In this chapter, we propose a comprehensive framework for studying the properties of Conditional Permutation Importance (*CPI*) in biomedical applications alongside tools for benchmarking variable importance estimators:

- Building on the previous literature on *CPI*, we develop theoretical results for the limitations regarding Permutation Importance (*PI*) and advantages of conditional Permutation Importance (*CPI*) given correlated inputs (section 5.1).

- We propose a novel implementation for *CPI* allowing us to combine the potential advantages of highly expressive base learners for prediction (a deep neural network) and a comparably lean Random Forest model as a conditional probability learner (section 5.2).

- We conduct extensive benchmarks on synthetic and heterogeneous multimodal real-world biomedical data tapping into different correlation levels and data-generating scenarios for both classification and regression (section 5.3).

5.1 . Permutation importance and its limitations

5.1.1 . The *permutation* approach leads to false detections in the presence of correlations

A known problem with *permutation* variable importance is that if features are correlated, their importance is typically over-estimated [Strobl et al., 2008], leading to a loss of type-I error control. However, this loss has not been precisely characterized yet, which we will work through for the linear case. We use the setting of [Mi et al., 2021], where the estimator $\hat{\mu}$, computed with empirical risk minimization under the training set, is used to assess variable importance on a new set of data (test set). We consider a regression model with a least-square loss function for simplicity. The importance of variable $\mathbf{x^j}$ is computed as follows:

$$\hat{m}^j = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left((y_i - \hat{\mu}(\mathbf{x_i^{(j)}}))^2 - (y_i - \hat{\mu}(\mathbf{x_i}))^2 \right). \tag{5.1}$$

Let $\varepsilon_i = y_i - \mu(\mathbf{x_i})$ for $i \in [\![n_{test}]\!]$. Re-arranging terms yields

$$\hat{m}^j = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\hat{\mu}(\mathbf{x_i}) - \hat{\mu}(\mathbf{x_i^{(j)}}))(2\mu(\mathbf{x_i}) - \hat{\mu}(\mathbf{x_i}) - \hat{\mu}(\mathbf{x_i^{(j)}}) + 2\varepsilon_i). \tag{5.2}$$

Mi et al. [2021] argued that these terms vanish when $n_{test} \to \infty$. But it is not the case as long as the training set is fixed. In order to get tractable computation, we assume that μ and $\hat{\mu}$ are linear functions: $\mu(\mathbf{x}) = \mathbf{xw}$ and $\hat{\mu}(\mathbf{x}) = \mathbf{x}\hat{\mathbf{w}}$. Let us further consider that $\mathbf{x^j}$ is a null feature, i.e. $w^j = 0$. This yields $\mathbf{xw} = x^j w^j + \mathbf{x^{-j}w^{-j}} = \mathbf{x^{-j}w^{-j}}$ which leads to

$$\begin{aligned}
\hat{m}^j =& \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\hat{\mu}(\mathbf{x_i}) - \hat{\mu}(\mathbf{x_i^{(j)}}))(2\mu(\mathbf{x_i}) - \hat{\mu}(\mathbf{x_i}) - \hat{\mu}(\mathbf{x_i^{(j)}}) + 2\varepsilon_i) \qquad (5.2) \\
=& \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\cancel{\mathbf{x_i^{-j}\hat{w}^{-j}}} + x_i^j \hat{w}^j - \cancel{\mathbf{x_i^{-j}\hat{w}^{-j}}} - \{x_i^j\}^\pi \hat{w}^j)(2\mathbf{x_i^{-j}w^{-j}} - 2\mathbf{x_i^{-j}\hat{w}^{-j}} \\
& - (x_i^j \hat{w}^j + \{x_i^j\}^\pi \hat{w}^j) + 2\varepsilon_i) \\
=& \frac{2\hat{w}^j}{n_{test}} \sum_{i=1}^{n_{test}} (x_i^j - \{x_i^j\}^\pi)(\mathbf{x_i^{-j}}(\mathbf{w^{-j}} - \mathbf{\hat{w}^{-j}}) + \varepsilon_i) - \cancel{\hat{w}^j((x_i^j)^2 - (\{x_i^j\}^\pi)^2)} \\
=& \frac{2\hat{w}^j}{n_{test}} \sum_{i=1}^{n_{test}} (x_i^j - \{x_i^j\}^\pi)(\mathbf{x_i^{-j}}(\mathbf{w^{-j}} - \mathbf{\hat{w}^{-j}}) + \varepsilon_i) \\
=& \frac{2\hat{w}^j}{n_{test}} \left\langle \mathbf{x^j} - \{\mathbf{x^j}\}^\pi, \mathbf{X^{-j}}(\mathbf{w^{-j}} - \mathbf{\hat{w}^{-j}}) + \boldsymbol{\varepsilon} \right\rangle
\end{aligned}$$

as $(\|\mathbf{x^j}\|^2 - \|\{\mathbf{x^j}\}^\pi\|^2) = 0$. Next, $\frac{1}{n_{test}} \langle \{\mathbf{x^j}\}^\pi, \mathbf{X^{-j}}(\mathbf{w^{-j}} - \mathbf{\hat{w}^{-j}}) \rangle \to 0$ and $\frac{1}{n_{test}} \langle \mathbf{x^j} - \{\mathbf{x^j}\}^\pi, \boldsymbol{\varepsilon} \rangle \to 0$ when $n_{test} \to \infty$ with speed $\frac{1}{\sqrt{n_{test}}}$ from the Berry-Essen theorem, assuming that the first three moments of these quantities are bounded and that the test samples are i.i.d. Let us assume that the correlation within $\mathbf{X}$ takes the following form: $\mathbf{x^j} = \mathbf{X^{-j}u} + \boldsymbol{\delta}$, where $\mathbf{u} \in \mathbb{R}^{p-1}$ and $\boldsymbol{\delta}$ is a random vector independent of $\mathbf{X^{-j}}$. By contrast, $\frac{2\hat{w}^j}{n_{test}} \langle \mathbf{x^j}, \mathbf{X^{-j}}(\mathbf{w^{-j}} - \mathbf{\hat{w}^{-j}}) \rangle$ has a non-zero limit $2\hat{w}^j \mathbf{u}^T Cov(\mathbf{X^{-j}})(\mathbf{w^{-j}} - \mathbf{\hat{w}^{-j}})$, where $Cov(\mathbf{X^{-j}}) = \lim_{n_{test} \to \infty} \frac{\mathbf{X^{-j}}^T \mathbf{X^{-j}}}{n_{test}}$ (remember that both $\mathbf{w^{-j}}$ and $\mathbf{\hat{w}^{-j}}$ are fixed, because the training set is fixed). Thus, the permutation importance of a null but correlated variable does not vanish when $n_{test} \to \infty$, implying that this inference scheme will lead to false positives.

5.2 . *Conditional sampling*-based feature importance

5.2.1 . Main result

We define the permutation of variable x^j conditional to $\mathbf{x}^{-\mathbf{j}}$, as a variable $\tilde{x}^j$ that retains the dependency of x^j with respect to the other variables in $\mathbf{x}^{-\mathbf{j}}$, but where the independent part is shuffled; $\tilde{\mathbf{x}}^{(\mathbf{j})}$ is the vector $\mathbf{x}$ where x^j is replaced by $\tilde{x}^j$. We propose two constructions below (see Fig. 5.1) that we compare in the additional experiments (section 5.5.4), one of which is faster than the other. In the case of regression, this leads to the following importance estimator:

$$\hat{m}^j_{CPI} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left((y_i - \hat{\mu}(\tilde{\mathbf{x}}^{(\mathbf{j})}_{\mathbf{i}}))^2 - (y_i - \hat{\mu}(\mathbf{x_i}))^2 \right). \tag{5.3}$$

As noted by Watson and Wright [2021], this inference is correct, as in tradi-

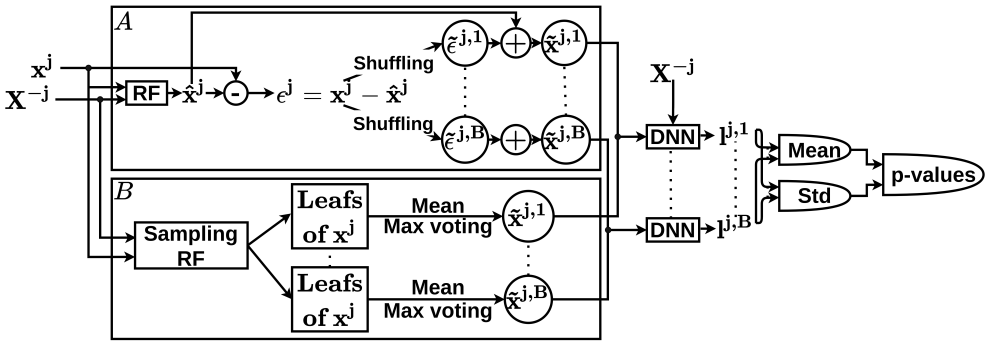

Figure 5.1: **Construction of CPI**: Constructing the variable of interest $\tilde{\mathbf{x}}^{\mathbf{j}}$ is done either (***A***) by the additive construction where a shuffled version of the residuals is added to the predicted version using the remaining predictors with the mean of a Random Forest (RF) or (***B***) by the sampling construction using a Random Forest (RF) model to fit $\mathbf{x}^{\mathbf{j}}$ from $\mathbf{X}^{-\mathbf{j}}$ and then sample the prediction within the leaves of the RF.

tional permutation tests, as long as one wishes to perform inference conditional to $\hat{\mu}$. However, the following proposition states that this inference has much wider validity in the asymptotic regime.

Proposition. *Assuming that the estimator $\hat{\mu}$ is obtained from a class of functions $\mathcal{F}$ with sufficient regularity, i.e. that it meets conditions (A1, A2, A3, A4, B1 and B2) defined in additional proofs, the importance score $\hat{m}^j_{CPI}$ defined in (5.3) cancels when $n_{train} \to \infty$ and $n_{test} \to \infty$ under the null hypothesis, i.e. the j^{th} variable is not significant for the prediction. Moreover, the Wald statistic $z^j = \frac{mean(\hat{m}^j_{CPI})}{std(\hat{m}^j_{CPI})}$ obtained by dividing the mean of the importance score by its standard deviation asymptotically follows a standard normal distribution.*

This implies that in the large sample limit, the p-value associated with z^j controls the type-I error rate for all optimal estimators in $\mathcal{F}$.

The proof of the proposition is given next (section 5.2.2). It consists in observing that the importance score defined in (5.3) is 0 for the class of learners discussed in [Williamson et al., 2021], namely those that meet a certain set of convergence guarantees and are invariant to arbitrary change of their j^{th} argument, conditional on the others. In the proof, we also restate the precise technical conditions under which the importance score $\hat{m}^j_{CPI}$ used is (asymptotically) valid, i.e. leads to a Wald-type statistic that behaves as a standard normal under the null hypothesis.

It is easy to see that for the setting in Sec. 5.1.1, all terms in Eq. 5.3 vanish with speed $\frac{1}{\sqrt{n_{test}}}$.

5.2.2 . Conditional Permutation Importance (CPI) Wald statistic asymptotically controls type-I errors: hypotheses, theorem and proof

Outline The proof relies on the observation that the importance score defined in (5.3) is 0 in the asymptotic regime, where the permutation procedure becomes a sampling step, under the assumption that variable j is not conditionally associated with y. Then all the proof focuses on the convergence of the finite-sample estimator to the population one. To study this, we use the framework developed in [Williamson et al., 2021]. Note that the major difference with respect to other contributions [Watson and Wright, 2021] is that the ensuing inference is no longer conditioned on the estimated learner $\hat{\mu}$. Next, we first restate the precise technical conditions under which the different importance scores considered are asymptotically valid, i.e. lead to a Wald-type statistic that behaves as a standard normal under the null hypothesis.

Notations Let $\mathcal{F}$ represent the class of functions from which a learner $\mu : \mathbf{x} \mapsto y$ is sought.

Let P_0 be the data-generating distribution and P_n is the empirical data distribution observed after drawing n samples (noted n_{train} in the previous section; in this section, we denote it n to simplify notations). The separation between train and test samples is actually only relevant to alleviate some technical conditions on the class of learners used. $\mathcal{M}$ is the general class of distributions from which $P_1, \ldots, P_n, P_0$ are drawn. $\mathcal{R} := \{c(P_1 - P_2) : c \in [0, \infty), P_1, P_2 \in \mathcal{M}\}$ is the space of finite signed measures generated by $\mathcal{M}$. Let l be the loss function used to obtain μ. Given $f \in \mathcal{F}$, $l(f; P_0) = \int l(f(\mathbf{x}), y) P_0(\mathbf{z}) d\mathbf{z}$, where $\mathbf{z} = (\mathbf{x}, y)$. Let μ_0 denote a population solution to the estimation problem $\mu_0 \in \text{argmin}_{f \in \mathcal{F}} l(f; P_0)$ and $\hat{\mu}_n$ a finite sample estimate $\hat{\mu}_n \in \text{argmin}_{f \in \mathcal{F}} l(f; P_n) = \frac{1}{n} \sum_{(\mathbf{x}, y) \in P_n} l(f(\mathbf{x}), y)$.

Let us denote by $\dot{l}(\mu, P_0; h)$ the Gâteaux derivative of $P \mapsto l(\mu, P)$ at P_0 in the direction $h \in \mathcal{R}$, and define the random function $g_n : \mathbf{z} \mapsto \dot{l}(\hat{\mu}_n, P_0; \delta_{\mathbf{z}} -$

$P_0) - \dot{l}(\mu_0, P_0; \delta_{\mathbf{z}} - P_0)$, where $\delta_{\mathbf{z}}$ is the degenerate distribution on $\mathbf{z} = (\mathbf{x}, y)$.

Hypotheses

(A1) (Optimality) there exists some constant $C > 0$, such that for each sequence $\mu_1, \mu_2, \cdots \in \mathcal{F}$ given that $\|\mu_n - \mu_0\| \to 0, |l(\mu_n, P_0) - l(\mu_0, P_0)| < C\|\mu_n - \mu_0\|_{\mathcal{F}}^2$ for each n large enough.

(A2) (Differentiability) there exists some constant $\kappa > 0$ such that for each sequence $\epsilon_1, \epsilon_2, \cdots \in \mathbb{R}$ and $h_1, h_2, \cdots \in \mathcal{R}$ satisfying $\epsilon_n \to 0$ and $\|h_n - h_\infty\| \to 0$, it holds that

$$\sup_{\mu \in \mathcal{F}: \|\mu - \mu_0\|_{\mathcal{F}} < \kappa} \left| \frac{l(\mu, P_0 + \epsilon_n h_n) - l(\mu, P_0)}{\epsilon_n} - \dot{l}(\mu, P_0; h_n) \right| \to 0.$$

(A3) (Continuity of optimization) $\|\mu_{P_0 + \epsilon h} - \mu_0\|_{\mathcal{F}} = O(\epsilon)$ for each $h \in \mathcal{R}$.

(A4) (Continuity of derivative) $\mu \mapsto \dot{l}(\mu, P_0; h)$ is continuous at μ_0 relative to $\|.\|_{\mathcal{F}}$ for each $h \in \mathcal{R}$.

(B1) (Minimum rate of convergence) $\|\hat{\mu}_n - \mu_0\|_{\mathcal{F}} = o_P(n^{-1/4})$.

(B2) (Weak consistency) $\int g_n(\mathbf{z})^2 dP_0(\mathbf{z}) = o_P(1)$.

(B3) (Limited complexity) there exists some P_0-Donsker class $\mathcal{G}_0$ such that $P_0(g_n \in \mathcal{G}_0) \to 1$.

Proposition (Theorem 1 in [Williamson et al., 2021]) If the above conditions hold, $l(\hat{\mu}_n, P_n)$ is an asymptotically linear estimator of $l(\mu_0, P_0)$ and $l(\hat{\mu}_n, P_n)$ is non-parametric efficient.

Let $P_0^\star$ be the distribution obtained by sampling the j^{th} coordinate of $\mathbf{x}$ from the conditional distribution of $q_0(x^j|\mathbf{x}^{-\mathbf{j}})$, obtained after marginalizing over y:

$$q_0(x^j|\mathbf{x}^{-\mathbf{j}}) = \frac{\int P_0(\mathbf{x}, y)dy}{\int P_0(\mathbf{x}, y)dx^j dy}$$

$P_0^\star(\mathbf{x}, y) = q_0(x^j|\mathbf{x}^{-\mathbf{j}}) \int P_0(\mathbf{x}, y)dx^j$. Similarly, let $P_n^\star$ denote its finite-sample counterpart. It turns out from the definition of $\hat{m}_{CPI}^j$ in Eq. 5.3 that $\hat{m}_{CPI}^j = l(\hat{\mu}_n, P_n^\star) - l(\hat{\mu}_n, P_n)$. It is thus the final-sample estimator of the population quantity $m_{CPI}^j = l(\hat{\mu}_0, P_0^\star) - l(\hat{\mu}_0, P_0)$.

Given that $\hat{m}_{CPI}^j = l(\hat{\mu}_n, P_n^\star) - l(\hat{\mu}_0, P_0^\star) - (l(\hat{\mu}_n, P_n) - l(\hat{\mu}_0, P_0)) + l(\hat{\mu}_0, P_0^\star) - l(\hat{\mu}_0, P_0)$, the estimator $\hat{m}_{CPI}^j$ is asymptotically linear and non-parametric efficient.

The crucial observation is that under the j-null hypothesis, y is independent of x^j given $\mathbf{x}^{-\mathbf{j}}$. Indeed, in that case $P_0(\mathbf{x}, y) = q_0(x^j|\mathbf{x}^{-\mathbf{j}})P_0(y|\mathbf{x}^{-\mathbf{j}})P_0(\mathbf{x}^{-\mathbf{j}})$ and $P_0(x^j|\mathbf{x}^{-\mathbf{j}}, y) = P_0(x^j|\mathbf{x}^{-\mathbf{j}})$, so that $P_0^\star = P_0$. Hence, mean/variance of $\hat{m}_{CPI}^j$'s distribution provide

valid confidence intervals for m^j_{CPI} and $mean(\hat{m}^j_{CPI}) \underset{n\to\infty}{\to} 0$. Thus, the Wald statistic $\hat{z}^j_{CPJ}$ defined in section (4.2) converges to a standard normal distribution, implying that the ensuing test is valid.

In practice, hypothesis (B3), which is likely violated, is avoided by the use of cross-fitting as discussed in [Williamson et al., 2021]: as stated in the main text, variable importance is evaluated on a set of samples not used for training. An interesting impact of the cross-fitting approach is that it reduces the hypotheses to (A1) and (A2), plus the following two:

(B'1) (Minimum rate of convergence) $\|\hat{\mu}_n - \mu_0\|_{\mathcal{F}} = o_P(n^{-1/4})$ on each fold of the sample splitting scheme.

(B'2) (Weak consistency) $\int g_n(\mathbf{z})^2 dP_0(\mathbf{z}) = o_P(1)$ on each fold of the sample splitting scheme.

5.2.3 . Practical estimation

Next, we present algorithms for computing conditional permutation importance. We propose two constructions for $\tilde{x}^j$, the conditionally permuted counterpart of x^j. The first one is additive: on test samples, x^j is divided into the predictable and random parts $\tilde{x}^j = \mathbb{E}(x^j|\mathbf{x^{-j}}) + \left(x^j - \mathbb{E}(x^j|\mathbf{x^{-j}})\right)^{\pi}$, where the residuals of the regression of x^j on $\mathbf{x^{-j}}$ are shuffled to obtain $\tilde{x}^j$. In practice, the expectation is obtained by a universal but efficient estimator, such as a random forest trained on the test set.

The other possibility consists in using a random forest (RF) model to fit x^j from $\mathbf{x^{-j}}$ and then sample the prediction within leaves of the RF.

Random shuffling is applied B times. For instance, using the additive construction, a shuffling of the residuals $\tilde{e}^{\mathbf{j,b}}$ for a given $b \in [\![B]\!]$ allows to reconstruct the variable of interest as the sum of the predicted version and the shuffled residuals, that is

$$\tilde{\mathbf{x}}^{\mathbf{j,b}} = \hat{\mathbf{x}}^{\mathbf{j}} + \tilde{e}^{\mathbf{j,b}}. \tag{5.4}$$

Let $\tilde{\mathbf{X}}^{\mathbf{j,b}} = (\mathbf{x^1}, \ldots, \mathbf{x^{j-1}}, \tilde{\mathbf{x}}^{\mathbf{j,b}}, \mathbf{x^{j+1}}, \ldots, \mathbf{x^p}) \in \mathbb{R}^{n_{test} \times p}$ be the new design matrix including the reconstructed version of the variable of interest $\mathbf{x^j}$. Both $\tilde{\mathbf{X}}^{\mathbf{j,b}}$ and the target vector $\mathbf{y}$ are fed to the loss function in order to compute a loss score $l^{j,b}_i \in \mathbb{R}$ defined by

$$l^{j,b}_i = \begin{cases} y_i \log\left(\frac{S(\hat{y}_i)}{S(\tilde{y}^b_i)}\right) + (1-y_i)\log\left(\frac{1-S(\hat{y}_i)}{1-S(\tilde{y}^b_i)}\right) \\ (y_i - \tilde{y}^b_i)^2 - (y_i - \hat{y}_i)^2 \end{cases} \tag{5.5}$$

for binary and regression cases respectively where $i \in [\![n_{test}]\!]$, $j \in [\![p]\!]$, $b \in [\![B]\!]$, i indexes a test sample of the dataset, $\hat{y}_i = \hat{\mu}(\mathbf{x_i})$ and $\tilde{y}^b_i = \hat{\mu}(\tilde{\mathbf{x}}^{\mathbf{j,b}}_{\mathbf{i}})$ is the new fitted value following the reconstruction of the variable of interest with the b^{th} residual shuffled and S(x) = $\frac{1}{1+e^{-x}}$.

The variable importance scores are computed as the double average over the number of permutations B and the number of test samples n_{test} (line 15

of Alg. 1), while their standard deviations are computed as the square root of the average over the test samples of the quadratic deviation over the number of permutations (line 16). Note that, unlike Williamson et al. [2021], the variance estimator is non-vanishing, and thus can be used as a plugin. A z^j_{CPI} statistic is then computed by dividing the mean of the corresponding importance scores with the corresponding standard deviation (line 17). P-values are computed using the cumulative distribution function of the standard normal distribution (line 18). The conditional sampling and inference steps are summarized in Algorithm 1. This leads to the *CPI-DNN* method when $\hat{\mu}$ is a deep neural network, or *CPI-RF* when $\hat{\mu}$ is a random forest. Supplementary analysis reporting the computational advantage of *CPI-DNN* over a remove-and-relearn alternative a.k.a. *LOCO-DNN*, can be found in additional experiments (section 5.5.1), which justifies its *computational leanness*.

Algorithm 1 Conditional sampling step: The algorithm implements the conditional sampling step in place of the permutation approach when computing the p-value of variable x^j

Input: $\mathbf{X} \in \mathbb{R}^{n_{test} \times p}$, $\mathbf{y} \in \mathbb{R}^{n_{test}}$, $\hat{\mu}$: estimator, l: loss function, RF_j: learner trained to predict x^j from $\mathbf{x}^{-\mathbf{j}}$

1 $B \leftarrow$ number of permutations
2 $\mathbf{X}^{-\mathbf{j}} \leftarrow \mathbf{X}$ with j^{th} column removed
3 **for** *i = 1 to* n_{test} **do**
4 $\quad \hat{x}^j_i \leftarrow \text{RF}_j(\mathbf{x}^{-\mathbf{j}}_\mathbf{i})$
5 **end**
6 Residuals $\epsilon^\mathbf{j} \leftarrow \mathbf{x}^\mathbf{j} - \hat{\mathbf{x}}^\mathbf{j}$
7 **for** *b = 1 to B* **do**
8 $\quad \tilde{\epsilon}^{\mathbf{j},\mathbf{b}} \leftarrow$ Random Shuffling($\epsilon^\mathbf{j}$)
9 $\quad \tilde{\mathbf{x}}^{\mathbf{j},\mathbf{b}} \leftarrow \hat{\mathbf{x}}^\mathbf{j} + \tilde{\epsilon}^{\mathbf{j},\mathbf{b}}$
10 $\quad$ **for** *i = 1 to* n_{test} **do**
11 $\quad\quad \tilde{y}^b_i \leftarrow \hat{\mu}(\tilde{\mathbf{x}}^{\mathbf{j},\mathbf{b}}_\mathbf{i})$
12 $\quad\quad$ compute $l^{j,b}_i$
13 $\quad$ **end**
14 **end**
15 $\text{mean}(\hat{m}^j_{CPI}) = \frac{1}{n_{test}} \frac{1}{B} \sum_{i=1}^{n_{test}} \sum_{b=1}^{B} l^{j,b}_i$
16 $\text{std}(\hat{m}^j_{CPI}) = \sqrt{\frac{1}{n_{test}-1} \sum_{i=1}^{n_{test}} \left(\frac{1}{B} \sum_{b=1}^{B} l^{j,b}_i - mean(\hat{m}^j_{CPI}) \right)^2}$
17 $z^j_{CPI} = \frac{\text{mean}(\hat{m}^j_{CPI})}{\text{std}(\hat{m}^j_{CPI})}$
18 $p^j \leftarrow 1 - cdf(z^j_{CPI})$

5.3 . Experiments & Results

5.3.1 . Experiment 1: Type-I error control and accuracy when increasing variable correlation

We compare the performance of *CPI-DNN* with that of *Permfit-DNN* by applying both methods across different correlation scenarios. The data $\{\mathbf{x_i}\}_{i=1}^{n}$ follow a Gaussian distribution with a prescribed covariance structure $\mathbf{\Sigma}$ i.e. $\mathbf{x_i} \sim \mathcal{N}(0, \mathbf{\Sigma}) \forall i \in [\![n]\!]$. We consider a block-designed covariance matrix $\mathbf{\Sigma}$ of 10 blocks with an equal correlation coefficient $\rho \in \{0, 0.2, 0.5, 0.8\}$ among the variables of each block. In this experiment, $p = 100$ and $n = 300$. The first variable of each of the first 5 blocks is chosen to predict the target y with the following model, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$:

$$y_i = x_i^1 + 2\log(1 + 2(x_i^{11})^2 + (x_i^{21} + 1)^2) + x_i^{31} x_i^{41} + \epsilon_i, \ \forall i \in [\![n]\!]$$

The AUC score, type-I error, power and computation time are presented in Fig. 5.2. Based on the AUC scores, *Permfit-DNN* and *CPI-DNN* showed virtually identical performance. However, *Permfit-DNN* lost type-I error control when correlation in $\mathbf{X}$ is increased, while *CPI-DNN* always controlled the type-I error at the targeted rate. Considering power, both methods *Permfit-DNN* and *CPI-DNN* have almost similar power. In high correlation regime, Permfit-DNN yields more detections, but it does not control type-I errors (Fig. 5.2). Regarding computation time, *CPI-DNN* is slightly more computationally expensive than *Permfit-DNN*.

5.3.2 . Experiment 2: Performance across different settings

In the second setup, we check if *CPI-DNN* and *Permfit-DNN* control the type-I error with an increasing total number of samples n. The data are generated as previously, with a correlation $\rho = 0.8$. We fix the number of variables p to 50 while the number of samples n increases from 100 to 1000 with a step size of 100. We use 5 different models to generate the outcome $\mathbf{y}$ from $\mathbf{X}$: *classification*, *Plain linear*, *Regression with ReLu*, *Interactions only* and *Main effects with interactions*.

Classification The signal $\mathbf{X}\boldsymbol{\beta}^{\boldsymbol{main}}$ is turned to binomial variables using the probit function Φ. $\boldsymbol{\beta}^{main}$ and $\boldsymbol{\beta}^{\boldsymbol{quad}}$ are the two vectors with different lengths of regression coefficients having only $n_{\text{signal}} = 20$ non-zero coefficients, the true model. $\boldsymbol{\beta}^{\boldsymbol{main}}$ is used with the main effects while $\boldsymbol{\beta}^{\boldsymbol{quad}}$ is involved with the interaction effects. Following [Janitza et al., 2018], the $\boldsymbol{\beta}$ values $\in \{\boldsymbol{\beta}^{\boldsymbol{main}}, \boldsymbol{\beta}^{\boldsymbol{quad}}\}$ are drawn i.i.d. from the set $\mathcal{B} = \{\pm 3, \pm 2, \pm 1, \pm 0.5\}$.

$$y_i \sim Binomial(\Phi(\boldsymbol{x_i}\boldsymbol{\beta}^{\boldsymbol{main}})), \ \forall i \in [\![n]\!]$$

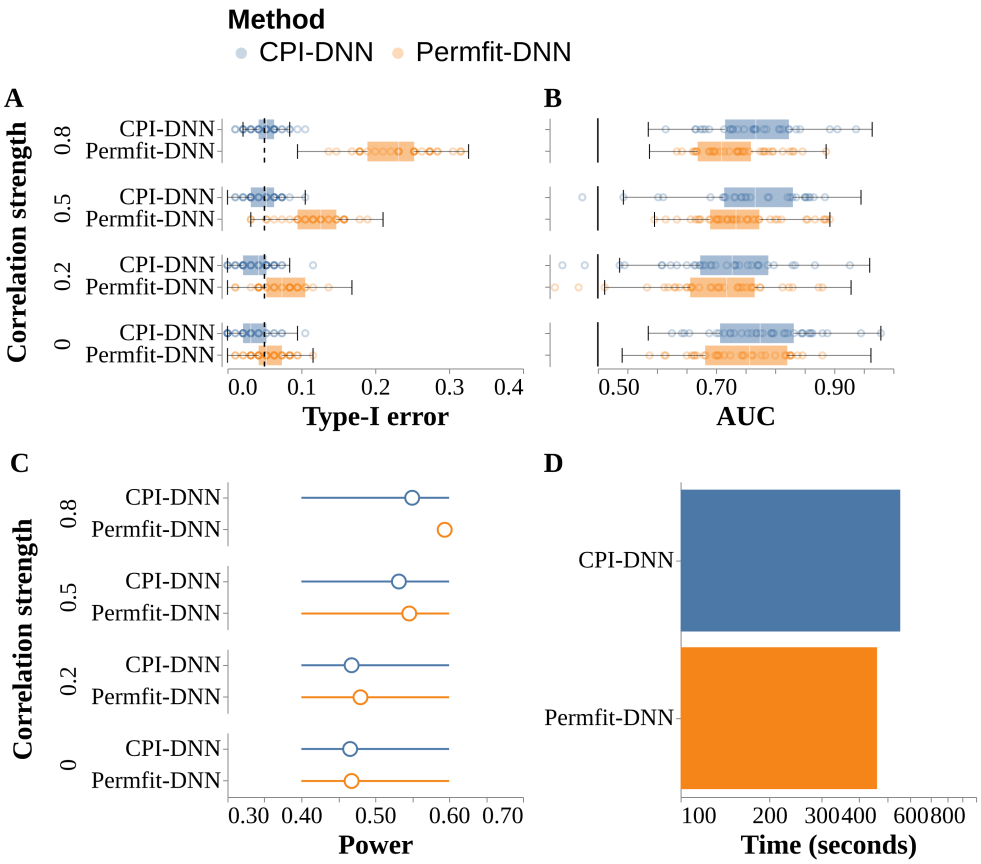

Figure 5.2: **CPI-DNN vs Permfit-DNN**: Performance at detecting important variables on simulated data with $n = 300$ and $p = 100$. **(A)**: The type-I error quantifies to which extent the rate of low p-values ($p < 0.05$) exceeds the nominal false positive rate. **(B)**: The AUC score measures to which extent variables are ranked consistently with the ground truth. **(C)**: The power reports the average proportion of informative variables detected (p-value < 0.05). **(D)**: The computation time is in seconds with (log10 scale) per core on 100 cores. Dashed line: targeted type-I error rate. Solid line: chance level.

Plain linear model We rely on a linear model, where $\boldsymbol{\beta}^{main}$ is drawn as previously and ϵ is the Gaussian additive noise $\sim \mathcal{N}(0, \mathbf{I})$ with magnitude $\sigma = \frac{||\mathbf{X}\boldsymbol{\beta}^{main}||_2}{SNR\sqrt{n}}$: $y_i = \mathbf{x_i}\boldsymbol{\beta}^{main} + \sigma\epsilon_i,\ \forall i \in [\![n]\!]$.

Regression with ReLu An extra ReLu function is applied to the output of the Plain linear model: $y_i = Relu(\mathbf{x_i}\boldsymbol{\beta^{main}} + \sigma\epsilon_i),\ \forall i \in [\![n]\!]$.

Interactions only model We compute the product of each pair of variables. The corresponding values are used as inputs to a linear model: $y_i = \text{quad}(\mathbf{x_i}, \boldsymbol{\beta^{quad}}) + \sigma\epsilon_i,\ \forall i \in [\![n]\!]$, where $\text{quad}(\boldsymbol{x_i}, \boldsymbol{\beta^{quad}}) = \sum_{\substack{k,j=1 \\ k<j}}^{p_{signals}} \boldsymbol{\beta}^{\boldsymbol{quad}}_{k,j} x_i^k x_i^j$.

The magnitude σ of the noise is set to $\frac{||\text{quad}(\boldsymbol{X},\boldsymbol{\beta^{quad}})||_2}{SNR\sqrt{n}}$. The non-zero $\boldsymbol{\beta^{quad}}$ coefficients are drawn uniformly from $\mathcal{B}$.

Main effects with Interactions We combine both Main and Interaction effects. The magnitude σ of the noise is set to $\frac{||\mathbf{X}\ \boldsymbol{beta}^{\text{main}} + \text{quad}(\boldsymbol{X},\boldsymbol{\beta^{quad}})||_2}{SNR\sqrt{n}}$: $y_i = \mathbf{x_i}\boldsymbol{\beta^{main}} + \text{quad}(\mathbf{x_i}, \beta^{\mathbf{quad}}) + \sigma\epsilon_i,\ \forall i \in [\![n]\!]$.

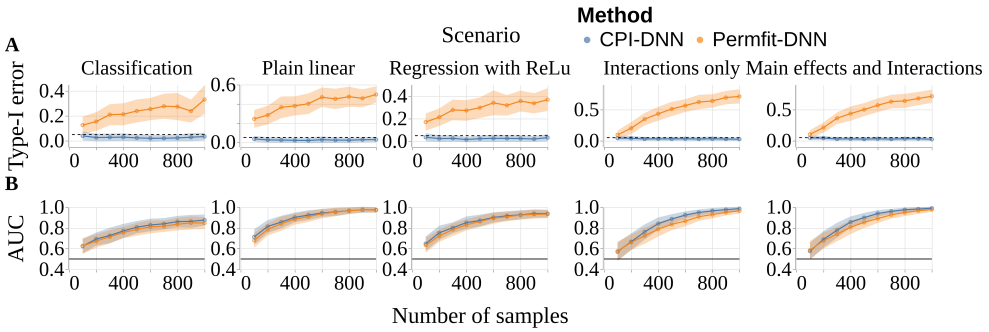

Figure 5.3: **Model comparisons across data-generating scenarios**: The **(A)** type-I error and **(B)** AUC scores of *Permfit-DNN* and *CPI-DNN* are plotted as function of sample size for five different settings. The number n of samples increased from 100 to 1000 with a step size of 100. The number of variables p was set to 50. Dashed line: targeted type-I error rate. Solid line: chance level.

The AUC score and type-I error of *Permfit-DNN* and *CPI-DNN* are shown as a function of sample size in Fig. 5.3. The accuracy of the two methods was similar across data-generating scenarios, with a slight reduction in the AUC scores of *Permfit-DNN* as compared to *CPI-DNN*. Only *CPI-DNN* controlled the rate of type-I error in the different scenarios at the specified level of 0.05. Thus, *CPI-DNN* provided an accurate ranking of the variables according to their importance score while, at the same time, controlling for the type-I error in all scenarios.

5.3.3 . Experiment 3: Performance benchmark across methods

In the third setup, we include *Permfit-DNN* and *CPI-DNN* in a benchmark with other state-of-the-art methods for variable importance using the same setting as in Experiment 2, while fixing the total number of samples n to 1000. We consider the following methods:

- Marginal Effects. A univariate linear model is fit to explain the response from each of the variables separately. The importance scores are then obtained from the ensuing p-values.
- Conditional-RF [Strobl et al., 2008]: A conditional variable importance approach based on a Random Forest model. This method provides p-values.
- d_0CRT [Liu et al., 2021, Nguyen et al., 2022]: The Conditional Randomization Test with distillation, using a sparse linear or logistic learner.
- Lazy VI [Gao et al., 2022].
- Permfit-DNN [Mi et al., 2021].
- LOCO [Lei et al., 2018]: This method applies the remove-and-retrain approach.
- cpi-knockoff [Watson and Wright, 2021]: Similar to CPI-RF, but permutation steps are replaced by a sampling step with a knockoff sampler.
- CPI-RF: This corresponds to the method in Alg. 1, where $\hat{\mu}$ is a Random Forest.
- CPI-DNN: This corresponds to the method in Alg. 1, where $\hat{\mu}$ is a DNN.

The extensive benchmarks on baselines and competing methods that provide p-values are presented in Fig. 5.4. For type-I error, *d_0CRT, CPI-RF*, *CPI-DNN*, *LOCO* and *cpi-knockoff* provided reliable control, whereas Marginal effects, *Permfit-DNN*, *Conditional-RF* and *Lazy VI* showed less consistent results across scenarios. For AUC, we observed that marginal effects performed poorly, as they do not use a proper predictive model. *LOCO* and *cpi-knockoff* behave similarly. *d_0CRT* performed well when the data-generating model was linear and did not include interaction effects. *Conditional-RF* and *CPI-RF* showed reasonable performance across scenarios. Finally, *Permfit-DNN* and *CPI-DNN* outperformed all the other methods, closely followed by *Lazy VI*. Additional inspection of power showed that across data generating scenarios, *CPI-DNN*, *Permfit-DNN* and *conditional-RF* showed strong results. *Marginal* and

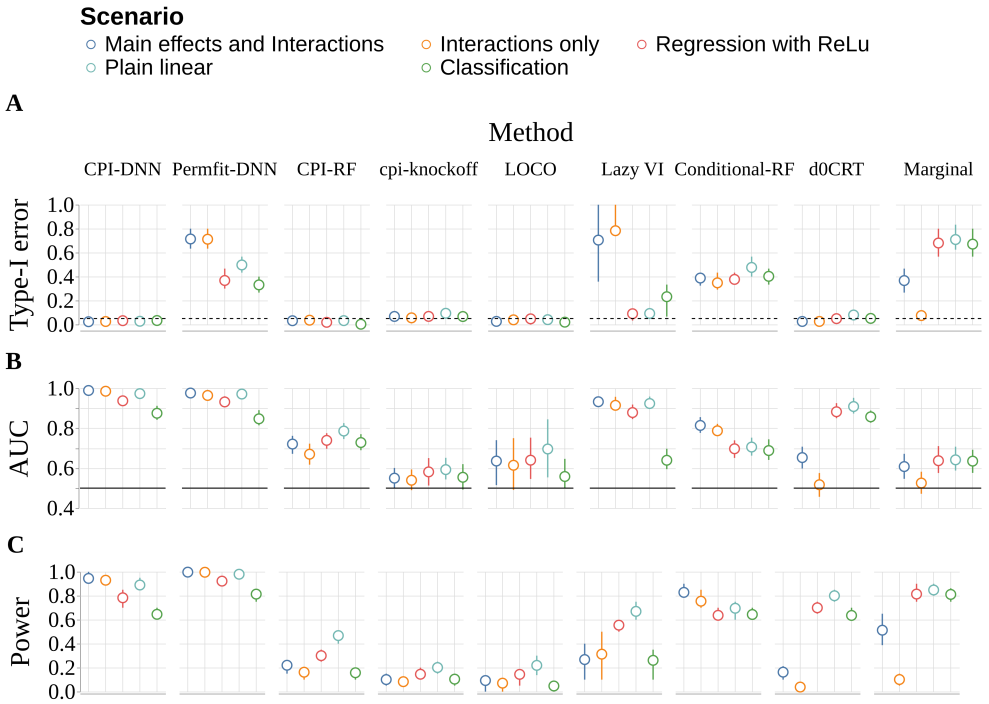

Figure 5.4: **Extended model comparisons**: *CPI-DNN* and *Permfit-DNN* were compared to baseline models (outer columns) and competing approaches across data-generating scenarios (inner columns). Prediction tasks were simulated with n = 1000 and p = 50. **(A)**: Type-I error. **(B)**: AUC scores. **(C)**: Power. Dashed line: targeted type-I error rate. Solid line: chance level.

d0CRT performed only well in scenarios without interaction effects. *CPI-RF*, *cpi-knockoff*, *LOCO* and *Lazy VI* performed poorly. To sum up, *Permfit-DNN* and *CPI-DNN* outperform the alternative methods. Thus, the use of the right learner leads to better interpretations.

We also benchmarked the following methods deprived of statistical guarantees i.e. not providing p-values:

- Knockoffs [Candes et al., 2017, Nguyen et al., 2020]
- Approximate Shapley values [Burzykowski, 2020].
- Shapley Additive Global importancE (SAGE) [Covert et al., 2020, Kumar et al., 2020].
- Mean Decrease of Impurity [Louppe et al., 2013].
- BART [Chipman et al., 2010].

The performance of these methods in terms of AUC score is reported in Fig. 5.5. Based on AUC, we observe SHAP, SAGE and Mean Decrease of

Impurity (MDI) perform poorly. These approaches are vulnerable to correlation. Next, Knockoff-Deep and Knockoff-Lasso perform well when the model does not include interaction effects. BART and Knockoff-Bart show fair performance overall.

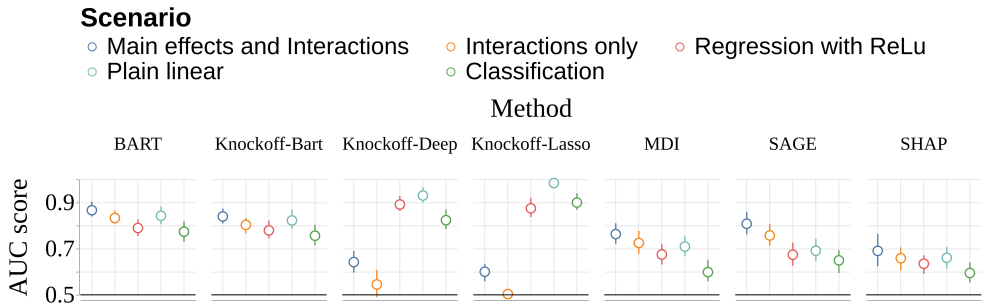

Figure 5.5: **Extended model comparisons-noPval**: State-of-the-art methods for variable importance not providing statistical guarantees in terms of p-values are compared (outer columns) and to competing approaches across data-generating scenarios (inner columns) using the settings of experiments 2 and 3. Prediction tasks were simulated with n = 1000 and p = 50. Solid line: chance level.

The computation time of the different methods mentioned in this work (with and without statistical guarantees) is presented in Fig. 5.6 in seconds with (log10 scale).

First, we compare *CPI-RF*, *cpi-knockoff* and *LOCO* based on a Random Forest learner with p=50. We see that *cpi-knockoff* and *LOCO* are faster than *CPI-DNN*. A possible reason is that *CPI-DNN* uses an inner 2-fold internal validation for hyperparameter tuning (learning rate, L1 and L2 regularization) unlike the alternatives. Next, The DNN-based methods (*CPI-DNN* and *Permfit-DNN*) are competitive with the alternatives that control type-I error (d_0CRT, *cpi-knockoff* and *LOCO*) despite the use of computationally lean learners in the latter.

Finally, to put estimated variable importance in perspective with model capacity, we benchmarked prediction performance of the underlying learning algorithms in Fig. 5.7, where the results for computing the prediction accuracy using the underlying learners of the different methods are reported. Marginal inference performs poorly, as it is not a predictive approach. Linear models based on Lasso show a good performance in the no-interaction effect scenario. Non-linear models based on Random Forest and BART improve on the lasso-based models. Nevertheless, they fail to achieve a good performance in scenarios with interaction effects. The models equipped with a deep learner outperform the other methods.

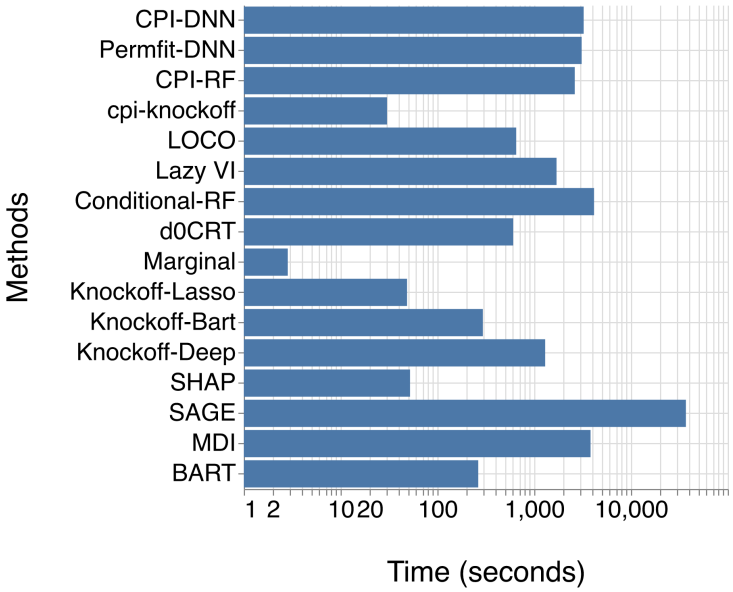

Figure 5.6: **Extended model comparisons**: The computation times for the different methods (with and without statistical guarantees in terms of p-values) are reported in seconds with (log10 scale) per core on 100 cores. Prediction tasks were simulated with n = 1000 and p = 50.

5.3.4 . Experiment 4: *Permfit-DNN* vs *CPI-DNN* on Real Dataset UKBB

Large-scale simulations comparing the performance of *CPI-DNN* and *Permfit-DNN* are conducted in additional experiments (section 5.5.2). We conducted an empirical study of variable importance in a biomedical application using the non-conditional permutation approach Permfit-DNN (no statistical guarantees for correlated inputs) and the safer CPI-DNN approach. A recent real-world data analysis of the UK Biobank dataset reported successful machine learning analysis of individual characteristics. The UK Biobank project (UKBB) curates phenotypic and imaging data from a prospective cohort of volunteers drawn from the general population of the UK [Constantinescu et al., 2022, Bycroft et al., 2018]. Nearly half a million people between the ages of 40 and 69 participated in the UK Biobank study, which began in 2006. These participants underwent various assessments, including physical tests, surveys about their background and lifestyle, cognitive tests, and medical examinations. The longitudinal study included two imaging visits where the number of participants dropped significantly to around 36 thousand and 8 thousand respectively. The data is provided by the UKBB operating within the terms of an Ethics and Governance Framework. The specifics of the processing pipeline are outlined below. The work focused on age, cognitive function and mood from brain images and social variables and put the ensuing models in rela-

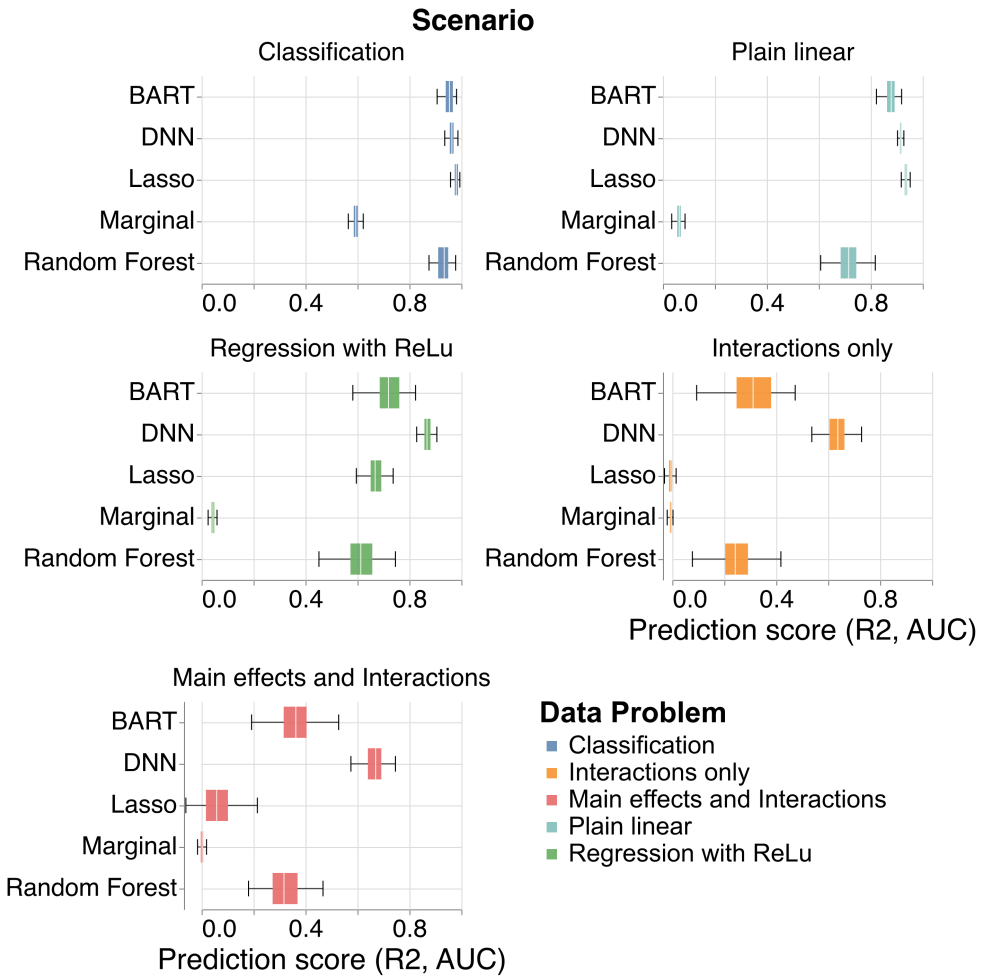

Figure 5.7: **Evaluation of predictive performance**: Performance of the different base learners used in the variable importance methods (**Marginal** = {Marginal effects}, **Lasso** = {Knockoff-Lasso}, **Random Forest** = {MDI, d0CRT, CPI-RF, Conditional-RF, cpi-knockoff, LOCO}, **BART** = {Knockoff-BART, BART} and **DNN** = {Knockoff-Deep, Permfit-DNN, CPI-DNN, Lazy VI}) on simulated data with $n = 1000$ and $p = 50$ in terms of **ROC-AUC** score for the classification and **R2** score for the regression.

tion to individual life-style choices regarding sleep, exercise, alcohol and tobacco [Dadi et al., 2021].

A coarse analysis of variable importance was presented, in which entire blocks of features were removed. It suggested that variables measuring brain structure or brain activity were less important for explaining the predictions of cognitive or mood outcomes than socio-demographic characteristics. On the other hand, brain imaging phenotypes were highly predictive of the age of a person, in line with the brain-age literature [Cole and Franke, 2017]. In

this benchmark, we explored variable-level importance rankings provided by the *CPI-DNN* and *Permfit-DNN* methods.

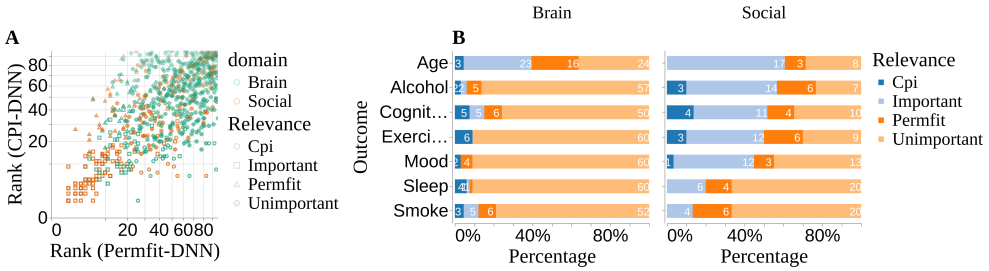

Figure 5.8: **Real-world empirical benchmark**: Prediction of personal characteristics (age, cognition, mood) and life-style habits (alcohol consumption, sleep, exercise & smoking) from various sociodemographic and brain-imaging derived phenotypes in a sample of $n = 8357$ volunteers from the UK Biobank. **(A)** plots variable rankings for *Permfit-DNN* (x axis) versus *CPI-DNN* (y axis) across all outcomes. Color: variable domain (brain versus social). Shape: variables classified by both methods as important (squares), unimportant (crosses) or by only one of the methods, *i.e.*, *CPI-DNN* (circles) or *Permfit-DNN* (triangles). **(B)** presents a detailed breakdown of percentage and counts of variable classification split by variable domain.

The real-world empirical benchmarks on predicting personal characteristics and life-style are summarized in Fig. 5.8. Results in panel **(A)** suggest that highest agreement for rankings between *CPI-DNN* and *Permfit-DNN* was achieved for social variables (bottom left, orange squares). At the same time, *CPI-DNN* flagged more brain-related variables as relevant (bottom right, circles). We next computed counts and percentage and broke down results by variable domain (Fig. 5.8, **B**). Naturally, the total relevance for brain versus social variables varied by outcome. However, as a tendency, *CPI-DNN* seemed more selective as it flagged fewer variables as important (blue) beyond those flagged as important by both methods (light blue). This was more pronounced for social variables where *CPI-DNN* sometimes added no further variables. As expected by the impact of aging on brain structure and function, brain data was most important for age-prediction compared to other outcomes. Interestingly, most disagreements between the methods occurred in this setting as *CPI* rejected 16 out of 66 brain inputs that were found as important by *Permfit*. This outlines the importance of correlations between brain variables, that lead to spurious importance findings with *Permfit*. We further explored the utility of our approach for age-prediction from neuromagnetic recordings [Engemann et al., 2020] and observed that *CPI-DNN* readily selected relevant frequency bands without fine-tuning the approach (Additional experiments section 5.5.3).

Processing pipelines

Structural MRI High-resolution brain volumes were extracted from T1-weighted MRI scans acquired using a Magnetization-Prepared Rapid Acquisition with Gradient Echo (MPRAGE) sequence at spatial resolution of 1x1x1 mm. Following de-identification, field distortion correction, reduction of Field of View (FoV), and skull-stripping using Brain Extraction Tool (BET) [Smith, 2002], the images were spatially normalized to MNI 152 T1 template space using non-linear registration method (FNIRT) [Andersson and Sotiropoulos, 2015] and segmented into most prominent tissue types such as gray matter, white matter, and cerebrospinal fluid volumes using FAST segmentation method [Zhang et al., 2001]. These bias-corrected images then underwent further processing to generate 157 Imaging Derived Phenotypes (IDPs) representing volumes of various cortical and subcortical structures, modeled using SIENAX [Smith et al., 2002] and FIRST [Patenaude et al., 2011] tools. We incorporated the 157 structural sMRI features, comprising the total brain volume, gray matter volume, and subcortical structures, into our analysis. These features were pre-extracted by the UKBB brain imaging team [Miller et al., 2016] and are included in the downloaded data. We utilize them in their original form, stacking them with other MRI features for predictive analysis.

Diffusion MRI Diffusion-weighted imaging, employing EPI sequences, was utilized for in-vivo measurement of local structures by tracking the movement of water molecules along fiber tracts. The dMRI data, acquired at a resolution of 2x2x2 mm with 50 diffusion-encoding directions and varying b-values (1000 and 2000), underwent preprocessing steps. These include correction for eddy current distortions, head motion, and removal of non-brain image slices [Andersson and Sotiropoulos, 2015]. Subsequently, gradient distortion correction is applied. The preprocessed images were then further processed to generate IDPs. This involved feeding the images into the Diffusion Tensor Imaging (DTIFIT) tool to model the diffusion directions, yielding IDPs such as Fractional Anisotropy (FA), Tensor Mode (MO), Mean Diffusivity (MD), and NODDI (Neurite Orientation Dispersion and Density Imaging) estimates utilizing Accelerated Microstructure Imaging via Convex Optimization (AMICO) [Daducci et al., 2015]. This process also allowed for modeling the biological properties of fiber tracts, represented as IDPs including Intra-Cellular Volume Fraction (ICVF), Isotropic Volume Fraction (ISOVF), and Orientation Dispersion index (OD).

To enable cross-subject comparisons on fiber tract-based IDPs, all outputs were aligned to a common space using tract-based spatial statistics (TBSS) [Smith et al., 2006]. We integrated 432 skeleton features derived from diffusion MRI (dMRI) data, encompassing Fractional Anisotropy (FA), Tensor Mode (MO), Mean Diffusivity (MD), as well as Intra-Cellular Volume Fraction (ICVF), Isotropic Volume Fraction (ISOVF), and Orientation Dispersion index

(OD). These features were modeled across various white matter structures.

rfMRI Resting-state functional magnetic resonance imaging (rfMRI) measures spontaneous low-frequency blood oxygen level-dependent (BOLD) signal fluctuations, which reflect ongoing interactions between large-scale brain networks [Biswal et al., 1995]. rfMRI data acquired using echo-planar imaging (EPI) sequences with multi-band acceleration was employed, yielding 490 brain volumes over a 6-minute scan. Preprocessing steps included motion correction, intensity normalization, high-pass filtering, EPI distortion correction, co-registration to a T1 template and grand-mean intensity normalization, and removal of structured artifacts via Independent Component Analysis (ICA) and FIX strategy. Subsequently, group-Principal Component Analysis (PCA) was applied for dimensionality reduction to facilitate the identification of high-resolution resting-state networks (RSNs) using ICA implemented in the MELODIC tool [Beckmann and Smith, 2004]. Artifactual components were excluded, and subject-specific time series signals were extracted through dual regression analysis. Regularized covariance was employed to estimate connectivity matrices, which were then mapped into a Euclidean space using tangent space embedding. Features for supervised learning were obtained by vectorizing the lower triangular portions of the connectivity matrices. Notably, Nilearn [Abraham et al., 2014] was utilized to implement the tangent space parametrization.

Resting-state connectivity features were incorporated based on the time-series derived from 55 ICA components representing diverse brain networks. Functional connectivity was assessed in terms of between-network covariance. In order to account for the fact that covariance matrices reside within a specific manifold, namely a curved non-Euclidean space, the tangent-space embedding was utilized to project the matrices into a Euclidean space [Varoquaux et al., 2010]. Subsequently, the connectivity matrices were transformed into a feature space of 1485 dimensions by vectorizing them, focusing on the lower triangular part for predictive modeling.

Socio-demographic data

This study builds upon the work of Dadi et al. [2021] and, in addition to brain scans, considers 86 pieces of non-imaging data. This data serves as a repository of information regarding each participant's background and social circumstances. Examples of the variables included in the study are age, sex, birthdate, body mass index, ethnicity, early life events (breastfeeding, maternal smoking, adoption), education level, lifestyle factors (occupation, income, household size, smoking habits), and any prior history of mental health conditions. The entire set of 86 variables is grouped into five clusters, with each cluster comprising variables that exhibit correlation associations with each other. These clusters are designated as follows: The first cluster comprises

age and sex, which are considered primary demographic variables. The second cluster encompasses early life experiences, while the third cluster encompasses education. The fourth cluster encompasses lifestyle factors, and the fifth cluster encompasses mental health.

5.4 . Discussion

In this work, we have developed a framework for studying the behavior of marginal and conditional permutation methods and proposed the *CPI-DNN* method, that was inspired by the limitations of the *Permfit-DNN* approach. Both methods build on top of an expressive DNN learner, and both methods turned out superior to competing methods at detecting relevant variables, leading to high AUC scores across various simulated scenarios. However, our theoretical results predicted that *Permfit-DNN* would not control type-I error with correlated data, which was precisely what our simulation-based analyzes confirmed for different data-generating scenarios (Fig. 5.2 - 5.3). Other popular methods (Fig. 5.4) showed similar failures of type-I error control across scenarios or only worked well in a subset of tasks. Instead, *CPI-DNN* achieved control of type-I errors by upgrading the *permutation* to *conditional permutation*. The consequences were pronounced for correlated predictive features arising from generative models with product terms, which was visible even with a small fraction of data points for model training. Among alternatives, the *Lazy VI* approach [Gao et al., 2022] obtained an accuracy almost as good as *Permfit-DNN* and *CPI-DNN* but with an unreliable type-I error control.

Taken together, our results suggest that *CPI-DNN* may be a practical default choice for variable importance estimation in predictive modeling. A practical validation of the standard normal distribution assumption for the non important variables can be found in additional experiments (section 5.5.5). The *CPI* approach is generic and can be implemented for any combination of learning algorithms as a base learner or conditional means estimator. *CPI-DNN* has a linear and quadratic complexity in the number of samples and variables, respectively. This is of concern when modeling the conditional distribution of the variable of interest which lends itself to high computational complexity. In our work, Random Forests proved to be useful default estimators as they are computationally lean and their model complexity, given reasonable default choices implemented in standard software, can be well controlled by tuning the tree depth. In fact, our supplementary analyses (section 5.5.6) suggest that proper hyperparameter tuning was sufficient to obtain good calibration of p-values. As a potential limitation, it is noteworthy the current configuration of our approach uses a deep neural network as the base learner. Therefore, in general, more samples might be needed for good model performance, hence, improved model interpretation.

Our real-world data analysis demonstrated that *CPI-DNN* is readily applicable, providing similar variable rankings as *Permfit-DNN*. The differences observed are hard to judge as the ground truth is not known in this setting. Moreover, accurate variable selection is important to obtain unbiased interpretations which are relevant for data-rich domains like econometrics, epidemiology, medicine, genetics or neuroscience. In that context, it is interesting that recent work raised doubts about the signal complexity in the UK biobank dataset [Schulz et al., 2020], which could mean that underlying predictive patterns are spread out over correlated variables. In the subset of the UK biobank that we analysed, most variables actually had low correlation values (Fig. 5.3.4-S2), which would explain why *CPI-DNN* and *Permfit-DNN* showed similar results. Nevertheless, our empirical results seem compatible with our theoretical results as *CPI-DNN* flagged fewer variables as important, pointing at stricter control of type-I errors, which is a welcome property for biomarker discovery.

When considering two highly correlated variables $\mathbf{x_1}$ and $\mathbf{x_2}$, the corresponding conditional importance of both variables is 0. This problem is linked to the very definition of conditional importance, and not to the *CPI* procedure itself. The only workaround is to eliminate, prior to importance analysis, degenerate cases where conditional importance cannot be defined. Therefore, possible future directions include inference on groups of variables, e.g, gene pathways, brain regions, while preserving statistical control offered by *CPI-DNN*.

5.5 . Additional Experiments

5.5.1 . Exp. 2 - Computational scaling of *CPI-DNN* and leanness

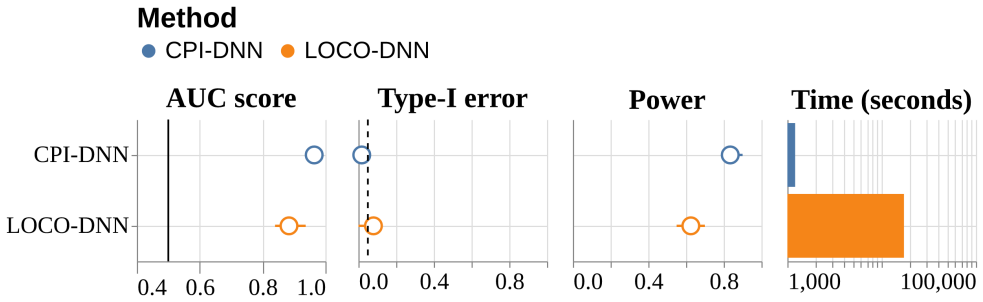

Figure 5.3.2-S1: ***CPI-DNN* vs *LOCO-DNN***: Performance at detecting important variables on simulated data with $n = 1000$, $p = 50$ and $\rho = 0.8$ in terms of **(AUC score)**, **Type-I error**, **Power** and **Time**. Dashed line: targeted type-I error rate. Solid line: chance level.

Computationally lean refers to two facts: (1) there is no need to refit the costly MLP learner to predict y unlike *LOCO-DNN* (A removal-based method provided with our learner) as seen in Fig. 5.3.2-S1. Both *CPI-DNN* and *LOCO-DNN* achieved a high AUC score and controlled the Type-I error in a highly correlated setting (ρ=0.8). However, in terms of computation time, *CPI-DNN* is far ahead of *LOCO-DNN*, which validates our use of the permutation scheme. (2) The conditional estimation step involved for the conditional permutation procedure is done with an efficient RF estimator, leading to small time difference wrt *Permfit-DNN*; Overall we obtain the accuracy of LOCO-type procedures for the cost of a basic permutation scheme.

5.5.2 . Exp. 4 - Large scale simulations

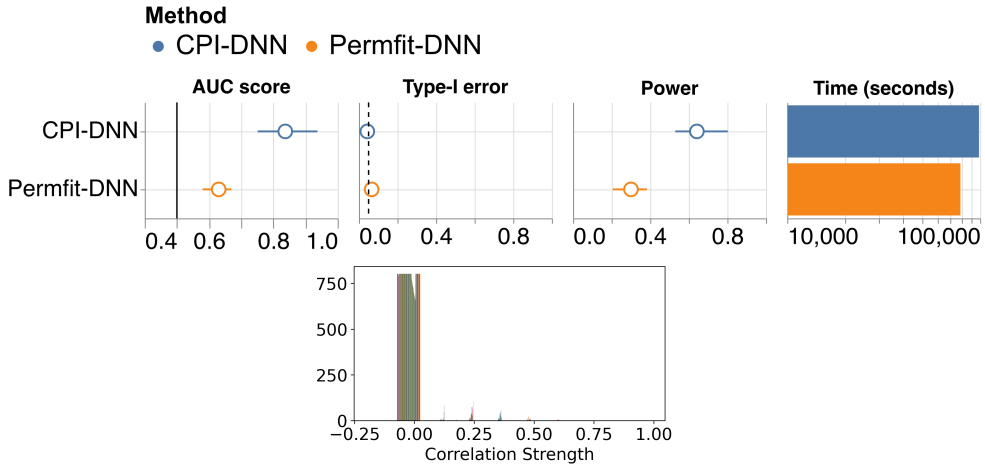

Figure 5.3.4-S1: **Semi-simulation with UK Biobank**: **(Top panel)** Performance of *CPI-DNN* and *Permfit-DNN* is compared in terms of **AUC score**, **Type-I error**, **Power** and **Time** on the data from UKBB with n = 8357 and p = 671. **(Bottom panel)** Correlation strength among the variables in the UKBB dataset.

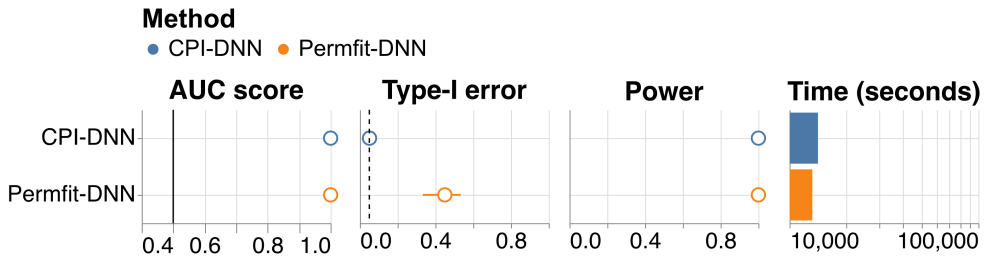

Figure 5.3.4-S2: **Large scale simulation**: Performance of *CPI-DNN* and *Permfit-DNN* is compared in terms of **AUC score**, **Type-I error**, **Power** and **Time** on simulated data with n = 10000, p = 50 and ρ = 0.8.

In Figs. 5.3.4-S1 and 5.3.4-S2, we provide a comparison of the performance of both *Permfit-DNN* and *CPI-DNN* on the semi-simulated data from UK Biobank, with the design matrix consisting of the variables in the UK BioBank and the outcome is generated following a random selection of the true support, where n=8357 and p=671, and a large scale simulation with n=10000, p = 50 and block-based correlation of coefficient $\rho = 0.8$. For the UKBB-based simulation, we see that *CPI-DNN* achieves a higher AUC score and Power. However, both methods control the type-I error at the targeted level. To better understand the reason, we plotted (Fig. 5.3.4-S1 Bottom panel) the histogram of the correlation values within the UKBB data: in this case, we consider a low-correlation setting which explains the good control for *Permfit-DNN*. In the large scale simulation where the correlation coefficient is set to 0.8, the difference is clear and only *CPI-DNN* controls the type-I error.

5.5.3 . Exp. 4 - Age prediction from brain activity (MEG) in Cam-CAN dataset

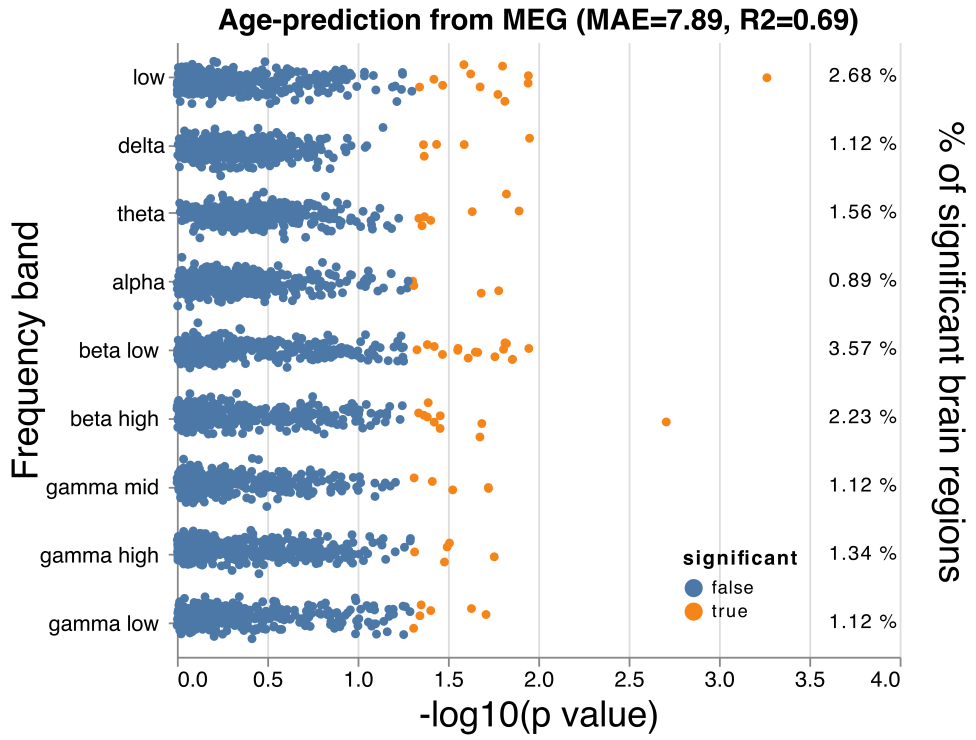

Figure 5-SE1: **Age prediction from brain activity**: Predicting age from brain activity in different frequencies with n = 536 and p = 4032.

Following the work of Engemann et al. [2020], we have applied *CPI-DNN* to the problem of age prediction from brain activity in different frequencies recorded with magnetoencephalography (MEG) in the Cam-CAN dataset.

Without tweaking, the DNN learner reached a prediction performance on par with the published results as seen in Fig. 5-SE1. The p-values formally confirm aspects of the exploratory analysis in the original publication (importance of beta band).

Dataset description

Launched in 2010, the Cambridge Center for Ageing and Neuroscience (Cam-CAN) [Taylor et al., 2017, Shafto et al., 2014] is a major collaborative research project investigating ageing and neuroscience. The study sought 3000 adults aged 18 and over to take part in a multi-stage manner. The first stage requested participants to participate in interviews about their health, lifestyle and life experiences, along with a cognitive assessment and a physical activity questionnaire. In the second stage, a subset of 700 participants, with 100 individuals representing each age group from 18 to 89, underwent additional cognitive assessments and brain imaging to measure both structural and functional characteristics. The cognitive tests assessed various domains, such as attention, memory, language, emotion and learning. This sample encompassed MEG (task and rest), fMRI (rest), anatomical MRI and neuropsychological data from 674 individuals (female = 340), aged between 18 (female = 18) to 88 (female = 87) under an average of 54.2 ± 18.7 (female = 53.7 ± 18.8) years.

Processing pipeline

MEG

MEG data was captured using a 306 VectorView system from Elekta Neuromag in Helsinki. This system facilitated the recording of magnetic fields employing 102 magnetometers and 204 orthogonal planar gradiometers within a lightly magnetically shielded room. Throughout the recording process, an online filter was implemented between 0.03Hz to 1000Hz. Following band-pass filtering between 0.1 and 49 Hz to isolate the neural signal of interest, the data underwent decimation by a factor of five. This resulted in a sampling frequency of 200 Hz during the subsequent epoching stage. To mitigate the influence of environmental magnetic noise on the MEG signal, the temporal signal space separation (tSSS) method was implemented [Taulu et al., 2005]. The analysis employed harmonic decomposition with default settings, incorporating eight components to capture internal sources and three components for external sources within a sliding window of ten seconds.

In order to ensure the quality of the data, data segments were excluded where the correlation between the inner and outer signal components fell below a threshold of 98%. Subsequently, a high-pass filter with a cutoff frequency of 0.1 Hz was applied to the signal, and the dimensionality of the data was reduced to approximately 65 dimensions. It is noteworthy that Maxwell filtering techniques, such as temporal signal space separation (tSSS), inte-

grate the signals from both magnetometers and gradiometers into a unified, low-rank representation. As a consequence of applying tSSS, the signal subsequently observed on magnetometers becomes a linear transformation of the signals initially observed on the gradiometers. This characteristic results in near-identical analytical outcomes when employing solely magnetometer data compared to gradiometer data [Garcés et al., 2017]. To optimize computational efficiency, the magnetometers were evaluated as a benchmark. To address the reduced data rank, a PCA projection was implemented to a shared rank of 65.

5.5.4 . Compare *CPI*'s constructions: *Residuals* vs *Sampling*

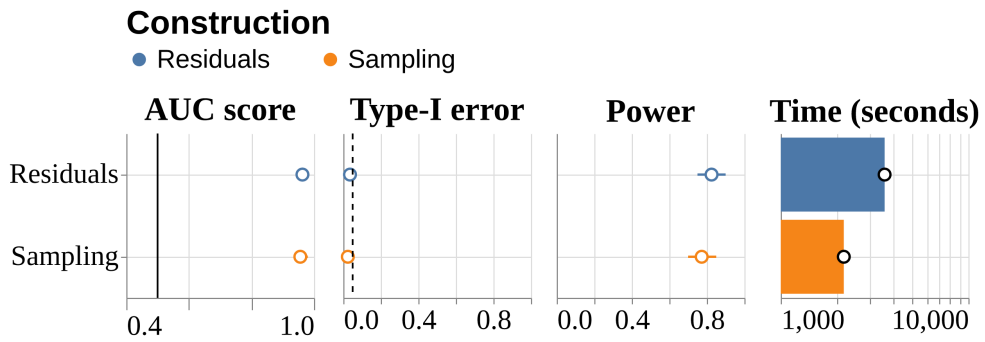

Figure 5-SE2: ***Residuals* vs *Sampling***: Comparison of *CPI*'s constructions, *Residuals* vs *Sampling*, in terms of **AUC score**, **Type-I error**, **Power** and **Time** on simulated data with n = 1000 and p = 50. Dashed line: targeted type-I error rate. Solid line: chance level.

5.5.5 . Practical validation of the normal distribution assumption

In Fig. 5-SE3, we compared the distribution of the importance scores of a random picked non-significant variable using *CPI-DNN* and *Permfit-DNN* through histogram plots, and we can emphasize that the normal distribution assumption holds in practice.

Also, in Fig. 5-SE4, we plot the distribution of the p-values provided by *CPI-DNN* and *Permfit-DNN* vs the uniform distribution through QQ-plot. We can see that the p-values for *CPI-DNN* are well calibrated and slightly deviated towards higher values. However, with *Permfit-DNN* the p-values are not calibrated.

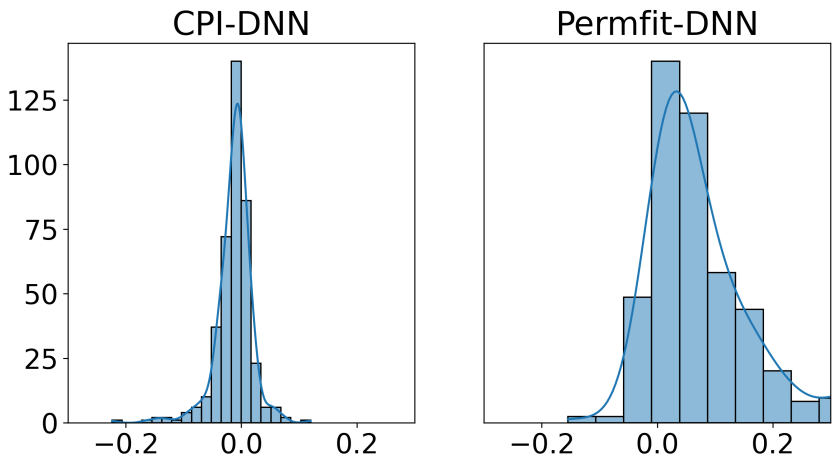

Figure 5-SE3: **Normal distribution assumption**: Histogram plots of the distribution of the importance scores of a random picked non-significant variable with $n = 1000$ and $p = 50$.

5.5.6 . Random Forest for modeling the conditional distribution and resulting calibration

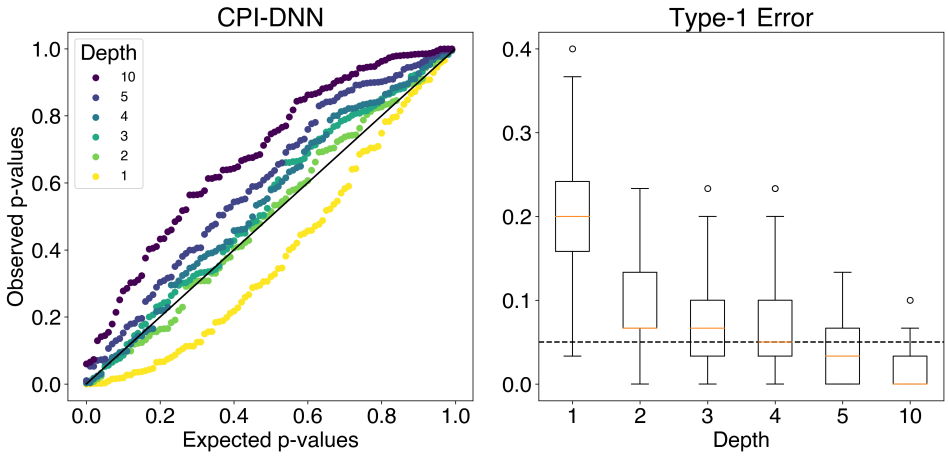

Figure 5-SE5: **Random forest calibration**: Calibration of the p-values for *CPI-DNN* (left panel) and the control of type-I error (right panel) as a function of the complexity of the Random Forest (the max depth of the trees). Dashed line: targeted type-I error rate. Solid line: uniform distribution.

The use of the Random Forest model was to maintain a good non-linear model with time benefits for the prediction of the conditional distribution of the variable of interest. In Fig. 5-SE5, We can see that reducing the depth to

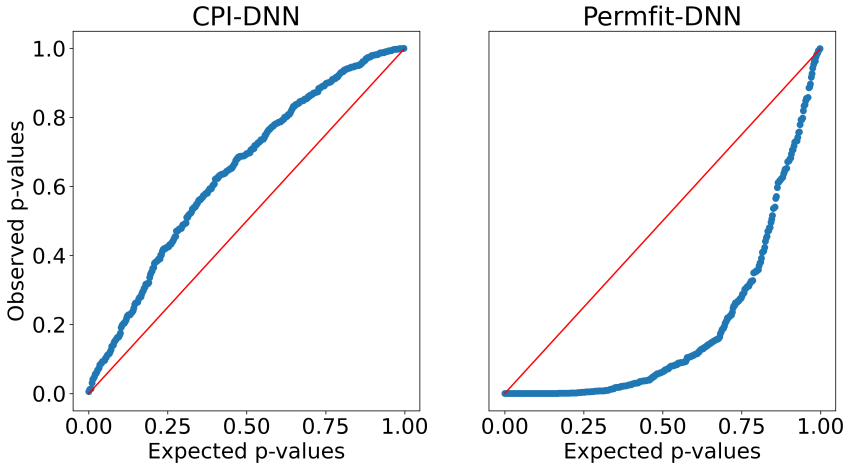

Figure 5-SE4: ***CPI-DNN* vs *Permfit-DNN* p-values calibration**: Q-Q plot for the distribution of the p-values vs the uniform distribution with n = 1000 and p = 50.

1 or 2, thus making the model overly simple, breaks the control of the type-I errors at the targeted level. With larger depths, the model becomes more conservative. Therefore, the max depth of the Random Forest is chosen based on the performance with 2-fold cross validation.

6 - Sampling of Continuous, Ordinal and Nominal Correlated Variables

6.1 . Background and Challenges

There is a growing interest in examining the influence of single variables on outcome prediction, a.k.a. *Variable Importance*, in real data applications across different domains. This requires utilizing a performance metric and ground truth to examine the validity of the deployed method. Nevertheless, a major challenge is the lack of necessary ground truth in real datasets, which forces us to consider simulations as an alternative. The variables in a dataset can be classified as either (1) *Continuous*, (2) *Categorical-ordinal* or (3) *Categorical-nominal*. The *Ordinal* and *Nominal* types are both defined by a limited number of values, with the former characterized by a natural order, such as rank vs color. Another challenge that arises is preserving the correlation between different variables. Therefore, a framework is needed to sample data with the same correlation among variables, meaning to simulate correlated variables with different types.

To perform the sampling procedure, the correlation matrix is a prerequisite. This matrix is computed using the *Ledoit-Wolf* estimator implemented in Scikit-learn [Pedregosa et al., 2011]. The sampling procedure started with the *Continuous* variables, followed by the *Categorical-ordinal* and finished with the *Categorical-nominal*. For the *Categorical* types, it is essential to maintain the marginal distributions along with the correlation structure.

6.2 . Sampling of *Continuous* and *Ordinal* variables

We focused on the sampling of *Continuous* and *Ordinal* variables, the former having a natural order as mentioned earlier. As stated in [Amatya and Demirtas, 2015], the generation of the *Ordinal* variables correlated with the *Continuous* variables went through *Normal* latent variables. Therefore, we aimed at this step to create

$$\#\textit{Normal} = \#\textit{Continuous} + \#\textit{Ordinal} \text{ variables.}$$

The procedure initiated by standard scaling the *Continuous* columns. Next, the generation of the targeted *Normal* variables was accomplished via the *Cholesky* decomposition. To illustrate, we considered one *Ordinal* variable $o_1 = \{0, 1, 2, 3\}$ and its corresponding *Normal* latent version l_1. The procedure proceeded by finding the *empirical cumulative distribution function* (F_{ecdf}) of o_1, i.e. $F_{ecdf}(o_1)$ and the *normal cumulative distribution function* (F_{norm}) of l_1, i.e.

$F_{norm}(l_1)$. To finish, under the equal distribution consideration, we searched for the corresponding category of the latent variable l_1.

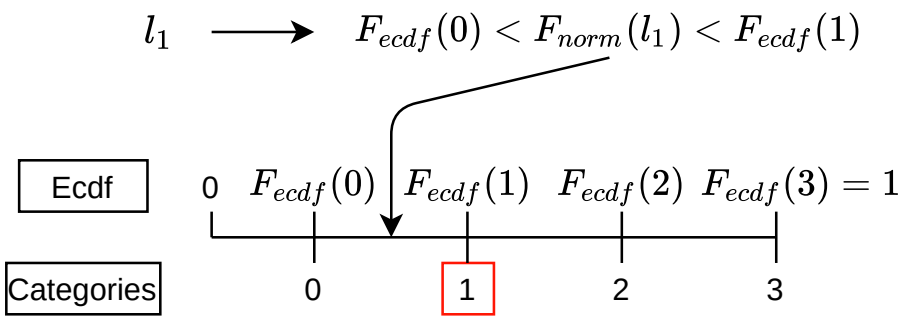

Figure 6.2.1: **Category specification of the latent variable with distribution equality**

In Fig. 6.2.1, we provided an example of the stated step where the latent variable l_1 was assigned the category (1).

6.3 . Sampling of *Nominal* variables

At this point, we assumed that the *Continuous* and *Ordinal* variables have already been sampled. The following step focused on the sampling of the *Nominal* variables distinguished by its iterative nature while considering the multiclass classification case. Hereafter, we used the *logistic regression* model for prediction purposes.

The *logistic regression* model fit the first *Nominal* variable using the *Continuous* and *Ordinal* variables and resampled a new copy of the variable. The resampled version was one-hot encoded and concatenated with the originally used variables (*Continuous* and *Ordinal*). Finally, the model moved to the next *Nominal* variable using the concatenated input variables.

Note that the sampling of *Binary* variables followed the same workflow as *Nominal* variables without the one-hot encoding step.

6.4 . Illustrative example

To provide an example for the whole procedure, we took the *UK Biobank* dataset [Constantinescu et al., 2022] where we extracted 10 *Continuous*, 2 *Ordinal* and 3 *Nominal* variables (one-hot encoded to 19 variables). We aimed to compare the computed correlation matrix under the *Ledoit-Wolf* method for the original and sampled variables. The results presented in Fig. 6.4.2-A demonstrated the effectiveness of the proposed regeneration procedure. To better illustrate the results, we applied the cross-correlation between the original and sampled observations in Fig. 6.4.2-B. The outcome was a noisy matrix, highlighting the fact that the observations were different after *sampling* while preserving the correlation structure.

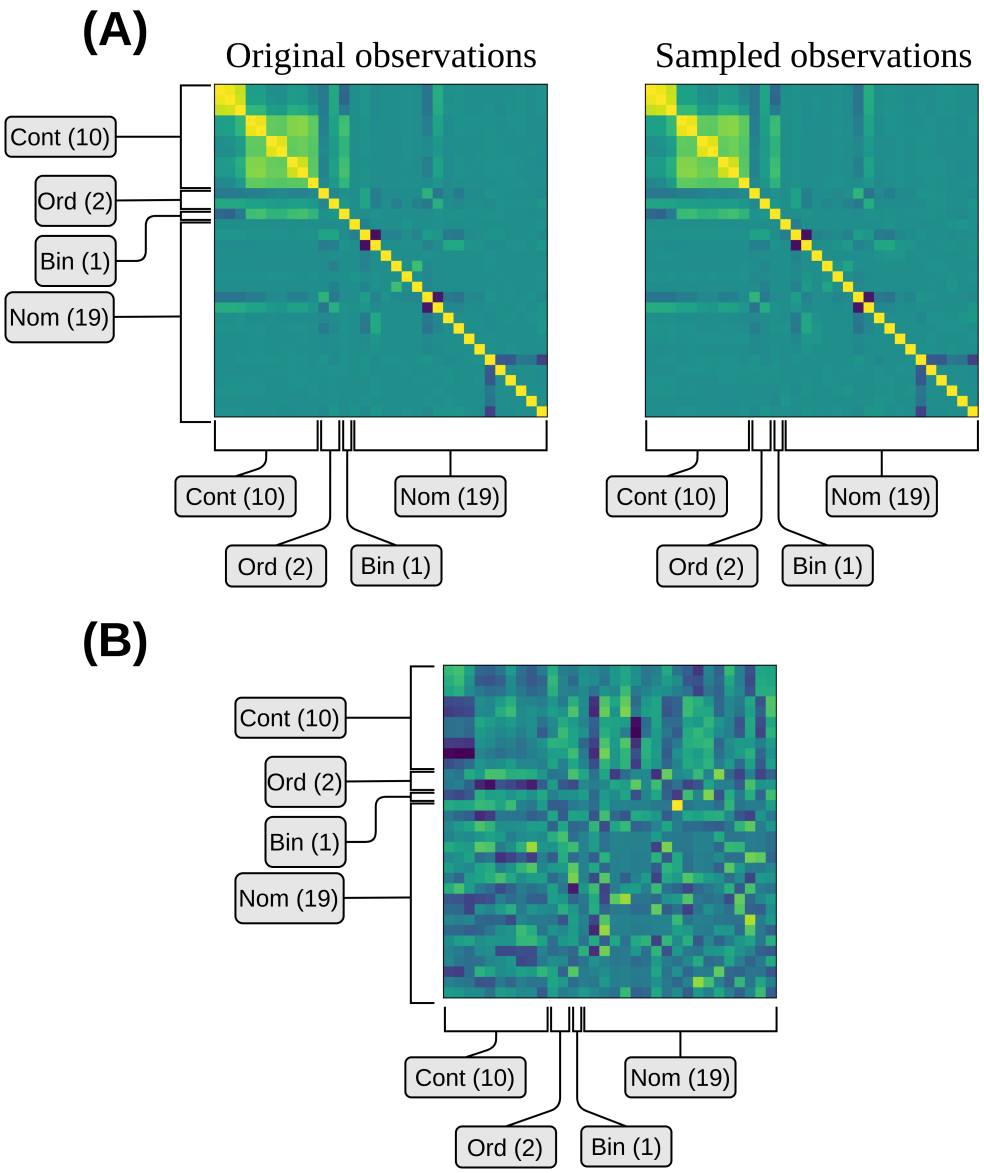

Figure 6.4.2: **Correlation-Adjusted Sampling**: Comparsion of the correlation structures after the sampling via the aforementioned procedure. **(A)** Correlation matrices of the original and sampled observations respectively from UKBB using *Ledoit-Wolf*. **(B)** Cross-correlation of the original and sampled observations.

7 - Statistical Valid Importance: the Case of Grouped Variables

Summary In this chapter, we propose *Block-Based Conditional Permutation Importance (BCPI)*, a new framework for variable importance computation (group levels) with explicit statistical guarantees (p-values).

- Following our review of the literature, we provide theoretical results on group-based conditional permutation importance (section 7.1).
- We propose a novel *internal stacking* approach by extending the architecture of our default Deep Neural Network (DNN) model with the use of a linear projection of the groups, which can significantly reduce computation time (section 7.2).
- We conduct extensive benchmarks on synthetic and real world data (section 7.3) which demonstrate the capacity of the proposed method to combine high prediction performance with statistically valid identification of important groups of variables.

7.1 . Block-Based Conditional Permutation Importance (BCPI)

7.1.1 . Define more notations for the groups

Let $\mathcal{S}' = \{\mathcal{G}'^k, k \in [\![K]\!]\}$ be the set of K new subset of variables following linear projections with a set $\mathcal{P}$ of projection matrices, $\mathcal{J} \in (\mathcal{S} \cup \mathcal{S}')$. Projection matrices are meant to produce a group summary of the information. Let $\mathcal{P} = \{\mathbf{U_k}, k \in [\![K]\!]\}$ be the set of projection matrices $\mathbf{U_k} \in \mathbb{R}^{|\mathcal{G}^k| \times |\mathcal{G}'^k|}$. Let $\mathbf{X}'$ be the linearly projected version of $\mathbf{X}$ via $\mathcal{P}$ where $p' = \sum_{k=1}^{K} |\mathcal{G}'^k|$.

7.1.2 . Group conditional variable importance

We define the joint permutation of group $\mathbf{x}^{\mathcal{J}}$ conditional to $\mathbf{x}^{-\mathcal{J}}$, as a group $\tilde{\mathbf{x}}^{\mathcal{J}}$ that preserves the joint dependency of $\mathbf{x}^{\mathcal{J}}$ with respect to the other variables in $\mathbf{x}^{-\mathcal{J}}$, although the independent part is shuffled. The reconstruction of $\tilde{\mathbf{x}}^{\mathcal{J}}$ is done via two approaches, both, based on fast approximation with a lean model: (**1**) Additive construction combines the prediction of a Random Forest using the remaining groups and a shuffled version of the residuals i.e. $\tilde{\mathbf{x}'}^{\mathcal{J}} = \mathbb{E}(\mathbf{x}'^{\mathcal{J}}|\mathbf{x}'^{-\mathcal{J}}) + (\mathbf{x}'^{\mathcal{J}} - \mathbb{E}(\mathbf{x}'^{\mathcal{J}}|\mathbf{x}'^{-\mathcal{J}}))^{\pi}$ where the residuals of the regression of $\mathbf{x}'^{\mathcal{J}}$ on $\mathbf{x}'^{-\mathcal{J}}$ are shuffled. **(2)** Sampling construction uses a Random Forest model to fit $\mathbf{x}'^{\mathcal{J}}$ from $\mathbf{x}'^{-\mathcal{J}}$, followed by sampling the prediction from within its leaves. When dealing with regression, this results in the following

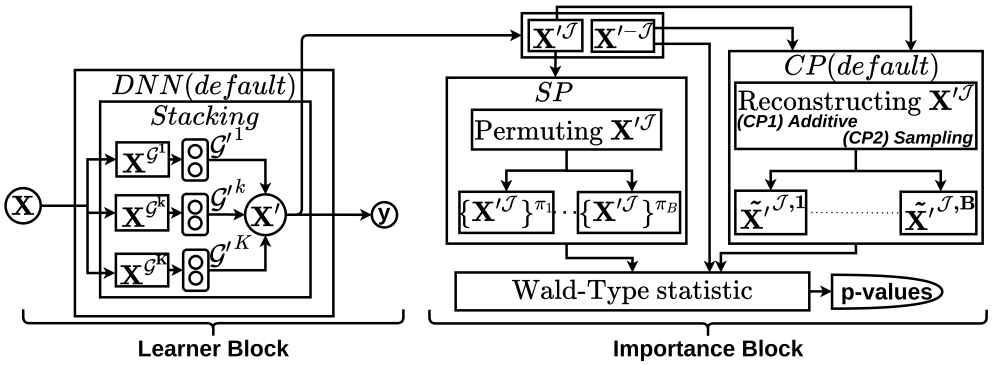

Figure 7.1.1: **Block-Based Conditional Permutation Importance**: Framework for single/group variable importance computation with statistical guarantees. **(Learner Block)** The learner used to predict the outcome $\mathbf{y}$ from the design matrix $\mathbf{X}$. *Internal stacking* linearly projects each group by the mean of an extra linear sub-layer. **(Importance Block)**: Reconstruction of the group of interest $\mathbf{X}'^{\mathcal{J}}$ is accomplished via *CP (Conditional Permutation)* block with (**CP1**) the additive or (**CP2**) the sampling constructions as stated in section 7.1.2. The permutation scheme can be changed to standard permutation (SP).

importance estimator:

$$\hat{m}_{CPI}^{\mathcal{J}} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left((y_i - \hat{\mu}(\tilde{\mathbf{x}}_{\mathbf{i}}^{(\mathcal{J})}))^2 - (y_i - \hat{\mu}(\mathbf{x}_{\mathbf{i}}))^2 \right), \tag{7.1}$$

where $\tilde{\mathbf{X}}^{(\mathcal{J})} = (\mathbf{x}^{\mathbf{1}}, \ldots, \mathbf{x}^{\mathbf{j_1-1}}, \tilde{\mathbf{x}}^{\mathbf{j_1}}, \ldots, \tilde{\mathbf{x}}^{\mathbf{j_r}}, \ldots, \mathbf{x}^{\mathbf{P}}) \in \mathbb{R}^{n_{test} \times p}$ be the new design matrix including the remodeled version of the group of interest $\mathbf{X}^{\mathcal{J}}$.

In Fig. 7.1.1, we introduce *BCPI* a novel general framework for variable importance, at both single and group levels, yielding statistically valid p-values. It consists of two blocks: a *Learner Block* defined by the prediction model of interest *Importance Block* reconstructing the variable (or group) of interest via conditional permutation (CP) – $\hat{m}_{CPI}^{\mathcal{J}}$. The implementation provided with this work supports estimators compatible with the scikit-learn API for both blocks. Yet, our default method *BCPI-DNN* is adapted with: (1) a DNN as a base learner for its high predictive capacity inspired from Mi et al. [2021] and (2) a Random Forest, a less powerful, but much simpler, yet, still generic model as a conditional probability learner. For study purposes, the framework is also adapted with the standard permutation scheme through the (SP) block (labeled *BPI*).

Proposition. *Assuming that the estimator $\hat{\mu}$ is obtained from a class of functions $\mathcal{F}$ with sufficient regularity, i.e. that it meets conditions of A1: optimality, A2: differentiability, A3: continuity of optimization, A4: Continuity of derivative, B1: Minimum rate of convergence and B2: Limited complexity, the importance score $\hat{m}_{CPI}^{\mathcal{J}}$*

defined in (7.1) cancels when $n_{train} \to \infty$ and $n_{test} \to \infty$ under the null hypothesis, i.e. the $\mathcal{J}^{th}$ group is not significant for the prediction. Moreover, the Wald statistic $z^{\mathcal{J}} = \frac{mean(\hat{m}^{\mathcal{J}}_{CPI})}{std(\hat{m}^{\mathcal{J}}_{CPI})}$ obtained by dividing the mean of the importance score by its standard deviation asymptotically follows a standard normal distribution.

The theoretical limitations of Permutation Importance (*PI*) have been explained in sections 5.1.1 considering a group of interest instead of a variable of interest. This implies that in the large sample limit, the p-value associated with $z^{\mathcal{J}}$ controls the type-I error rate for all optimal estimators in $\mathcal{F}$. The proof of the proposition is given in the additional proofs (section 5.2.2) with a group of interest. It entails making sure that the importance score defined in (7.1) is 0 for the class of learners that meet specific convergence guarantees and are immutable to arbitrary change in their $\mathcal{J}^{th}$ arguments, conditional on the others. We also state the precise technical conditions under with $\hat{m}^{\mathcal{J}}_{CPI}$ used is (asymptotically) valid, i.e. leads to a Wald-type statistic that behaves as a standard normal under the null hypothesis. As a result, all terms in Eq. 7.1 vanish with speed $\frac{1}{\sqrt{n_{test}}}$ from the *Berry-Essen* theorem, under the assumption that the test samples are i.i.d.

7.2 . *Internal* Stacking Approach

The vector $\mathbf{x} \in \mathcal{X}$ is composed of K groups in $\mathcal{S}$, each considered as an independent input modality. Performing column slicing on $\mathbf{x}$, according to $\mathcal{S}$, yields the set $\{\mathbf{x}^{\mathcal{G}^{\mathbf{k}}}, k \in [\![K]\!]\}$. A linear transformation to a lower space is applied on each input modality $\mathbf{x}^{\mathcal{G}^{\mathbf{k}}}$ through the set of projection matrices $\mathcal{P}$ producing a linear variant denoted $\mathbf{x'^{k}}$ as:

$$\mathbf{x'^{k}} =< \mathbf{x}^{\mathcal{G}^{\mathbf{k}}}, \mathbf{U_k} >,$$

where $k \in [\![K]\!]$.

Concatenating the set of linear variants $\{\mathbf{x'^{k}}, k \in [\![K]\!]\}$ provides the linearly projected version of $\mathbf{x}$ i.e. the vector $\mathbf{x'}$. If the new space is a one-dimensional Euclidean space i.e. $\mathbf{x'} \in \mathbb{R}^K$, a group summary of the information within all groups is returned, and the problem is reduced to the single-level case. However, if the new space is not unidimensional, we then have a dimension reduction, where the group summary of information is exclusive per group (multioutputs per group). In this case, the new groups contained in $\mathbf{x}$ are denoted $\mathcal{G'}^{k}$ with the corresponding linear variant $\mathbf{x'}^{\mathcal{G'}^{k}}$ as seen in Fig. 7.1.1. Instead of performing stacking in a separate estimation step under a different learner, we have incorporated it to the inference process, thus learning a consistent new presentation of the groups. This is simply implemented as an initial linear sub-layer without activation in the $\hat{\mu}$ network. Therefore, $\mathbf{x'^{k}}$ can be seen analogous to the predictions from the input

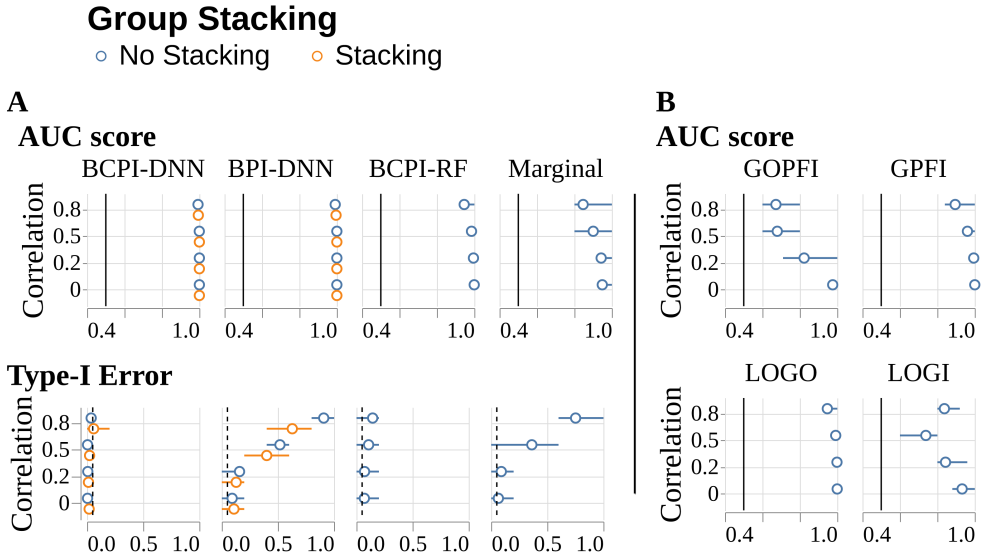

Figure 7.3.2: **Benchmarking grouping methods**: *BCPI-DNN* is compared to baseline models and competing approaches for group variable importance. **(A)** AUC score (correct ranking of variables) and Type-I error (p-val < 0.05) for methods providing p-values. **(B)** AUC scores for methods not providing p-values. Prediction tasks were simulated with n = 1000 and p = 50. Dashed line: targeted type-I error rate at 5%. Solid line: chance level.

models in a classical stacking pipeline that are forwarded to the meta learner, hence, $\mathbf{x}'^{\mathbf{k}}$ can be treated like a regular data column by permutation algorithms.

7.3 . Experiments

7.3.1 . Experiment 1: Benchmark of grouping methods

We include *BCPI-DNN* in a benchmark with other state-of-the-art methods for group-based variable importance. The data $\{\mathbf{x_i}\}_{i=1}^{n}$ follow a Gaussian distribution with a predefined covariance structure $\mathbf{\Sigma}$ i.e. $\mathbf{x_i} \sim \mathcal{N}(0, \mathbf{\Sigma}) \forall i \in [\![n]\!]$. We consider a block-designed covariance matrix $\mathbf{\Sigma}$ of 10 blocks with an intra-block correlation coefficient $\rho_{intra} = 0.8$ among the variables of each block and an inter-block correlation coefficient $\rho_{inter} \in \{0, 0.2, 0.5, 0.8\}$ between the variables of the different blocks. Each block is considered as a separate group. In this experiment, $n = 1000$ and $p = 50$ i.e. we have 5 variables per block/group. We defined an important group as a group having at least one variable that took part in simulating the outcome y. Thus, to predict y, we rely on a linear model where the first variable of each of the first 5 groups is used

in the following model:

$$y_i = \mathbf{x_i}\boldsymbol{\beta} + \sigma\epsilon_i, \forall i \in [\![n]\!] \tag{7.2}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients having only 5 non-zero coefficients (the true model), $\boldsymbol{\epsilon} \in \mathcal{N}(0, \mathbf{I})$ is the Gaussian additive noise with magnitude $\sigma = \frac{||\mathbf{X}\boldsymbol{\beta}||_2}{SNR\sqrt{n}}$. We used the same setting from Janitza et al. [2018] where the $\boldsymbol{\beta}$ values are drawn i.i.d. from the set $\mathcal{B} = \{\pm 3, \pm 2, \pm 1, \pm 0.5\}$. We consider the following state-of-the-art baselines:

- Marginal Effects: A multivariate linear model is applied to each group separately. Importance scores correspond to ensuing p-values.
- Leave-One-Group-In (*LOGI*) [Au et al., 2021]: Similar to *Marginal Effects* using a Random Forest. Provides no p-values.
- Leave-One-Group-Out (*LOGO*) [Williamson et al., 2021]: Refitting of the model is performed after removing the group of interest.
- Group Only Permutation Feature Importance (*GOPFI*) [Au et al., 2021]: Joint permutation of all variables except for those of the group of interest.
- Group Permutation Feature Importance (*GPFI*) [Gregorutti et al., 2015, Valentin et al., 2020]: Joint permutation of all variables of the group of interest.

In addition, we benchmarked the three variants of our proposed method:

- BPI-DNN: Similar to *GPFI* based on a DNN estimator. It is also reinforced by the new *internal stacking* approach.
- BCPI-RF: This corresponds to the method in Alg. 1, considering a group of interest, where $\hat{\mu}$ is a Random Forest.
- BCPI-DNN: This corresponds to the method in Alg. 1, considering a group of interest, where $\hat{\mu}$ is a DNN. It is also reinforced by the new *internal stacking* approach.

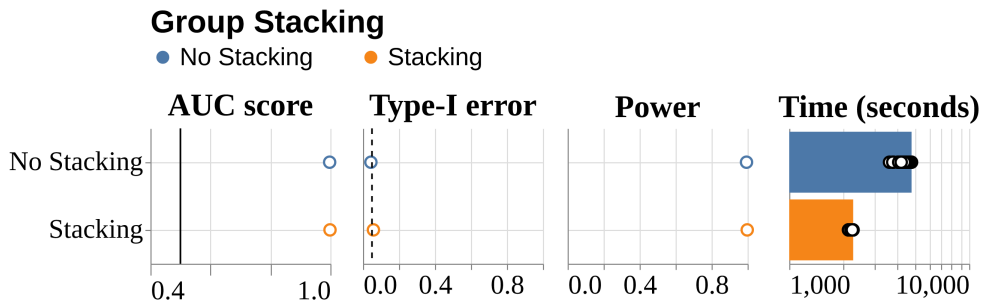

Figure 7.3.3: **Impact of Stacking**: Performance at detecting important groups on simulated data with n = 1000 and p = 1000 with 10 blocks/groups, each group having a cardinality of 10. AUC scores and Type-1 error as in Fig. 7.3.2. The **(Power)** provides information on the average proportion of detected informative variables (p-value < 0.05). The **(Time)** presents the time cost in seconds with log10 scale per core on 100 cores. Dashed line: targeted type-I error rate. Solid line: chance level.

7.3.2 . Experiment 2: Impact of Stacking

To assess the impact of performing stacking regarding accuracy in inference and computation time, we conducted a comparison restricted to *BCPI-DNN*. We relied on the same covariance structure setting as in Experiment 1 with an intra-block correlation coefficient $\rho_{intra} = 0.8$ and an inter-block correlation coefficient $\rho_{inter} = 0.8$. The number of samples n and the number of variables p were both set to 1000 i.e. the number of variables per block/group increased to 100 in order to build groups with high cardinality. The outcome y was simulated using the same model as in Eq. 7.2 where a group is predefined as important having at least 10% of its variables taking part in computing the outcome.

7.3.3 . Experiment 3: Age prediction with UKBB

We conducted an empirical benchmark of the performance of *BCPI-DNN* combined with *internal stacking* in a real-world biomedical dataset. The UK Biobank project (UKBB) encompasses imaging and socio-demographic derived phenotypes from a prospective cohort of participants drawn from the population of the UK [Constantinescu et al., 2022, Littlejohns et al., 2020]. In the past years, the UKBB dataset has enabled large-scale studies investigating associations between various phenotypes (physiological, cognitive) and environmental or life-style factor. This has given rise to successful analysis of factors associated to personal well-being and health [Newby et al., 2021, Mutz and Lewis, 2021] at an epidemiological scale. In the context of machine learning with brain data, age-prediction is an actively studied task which can provide a normative score when applying a reference model on clinical cohorts [Cole and Franke, 2017]. State-of-the-art models were based on con-

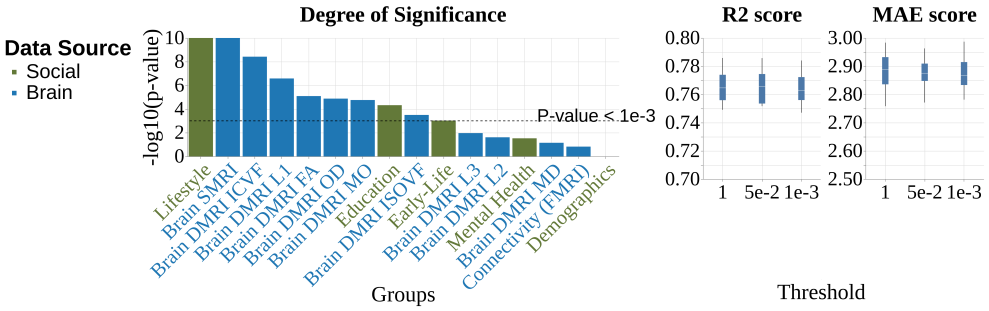

Figure 7.3.4: **Brain Age prediction in UKBB**: Prediction of brain age from various socio-demographic and brain-imaging groups of phenotypes in a sample of n = 8357 volunteers from the UK BioBank. (**Degree of significance**) plots the level of significance for the different brain (in blue) and social (in green) groups in terms of -log10 of the derived p-values. Dashed line: targeted type-I error rate at $p = 0.001$. (**R2 score & MAE score**) checks the predictive performance of the trained learner when retaining all the groups and after removing the non-important groups (having p-value $>$ 0.05 or 0.001).

volutional neural networks and report mean absolute errors between 2-3 years [Roibu et al., 2023, Jonsson et al., 2019]. Recent extensions have focused on MRI-contrast and region-specific insights, often based on informal inference [Roibu et al., 2023, Popescu et al., 2021]. Another line of work [Dadi et al., 2021, Anatürk et al., 2021] has focused on other sources of normative ageing information, highlighting cognitive social and lifestyle factors. In this context, the analysis of importance of multimodal inputs has so far been hampered by the lack of formal inference procedures and the high-dimensional setting with highly correlated variables.

We approached this open task using the proposed method, reusing the pre-defined groups in the work by Dadi et al. [2021] (supplement, section 7.7). We focused on data from participants who attended the imaging visit (n = 8357) to study the group-level importance rankings provided by *BCPI-DNN*. We defined important groups by p-value threshold of $< 10^{-3}$. While this setting lacks an explicit ground truth for the important groups, we explored the appropriate group selection through model performance in terms of (R^2 & MAE scores, 10-fold cross-validation) after removing the non-significant groups. We accessed the UKBB data through its controlled access scheme in accordance with its institutional ethics boards [Bycroft et al., 2018, Sudlow et al., 2015].

7.4 . Results

We benchmarked state-of-the-art baselines and the proposed methods across data-generating scenarios under increasing inter-block correlation strength {0, 0.2, 0.5, 0.8} (Fig. 7.3.2). *BCPI-DNN* and *BPI-DNN* were implemented in two variants: with or without the novel *internal stacking*. For the AUC score, we observed that (*BCPI-DNN* & *BPI-DNN* - based on the DNN) and (*BCPI-RF*, *GPFI* & *LOGO* - based on Random Forests) showed the highest performance across the different scenarios, hence, accurately ordering the variables according to their significance. As expected, the *Marginal* baseline performed lowest as it could not access any conditional information. *GOPFI* and *LOGI* both suffered when the correlation between the groups increased, which is not surprising. Considering false positive rate, *BCPI-DNN* controlled the type-I error at the targeted rate (5 %) while *BPI-DNN*— based on the standard permutation of the group of interest— failed to do so in the setting of high correlations between the groups, and thus provided spurious results. Interestingly, for *BPI-DNN*, *internal stacking* slightly increased its capacity to control the type-I error. *BCPI-RF*— based on the conditional importance with Random Forests— better controlled the type-I error compared to *BPI-DNN*. Nevertheless, in the presence of strong correlations, it did not fully reach the target rate. The supplement details the prediction performance for the different algorithms (section 7.6.1), suggesting that the *marginal* approach fails in the current setting, whereas on average, the DNN had higher scores ($R^2 \sim 0.95$) than the Random Forest ($R^2 \sim 0.8$). Performance in terms of power and computation time of these methods is reported in the additional experiments (section 7.6.1). The results showed that *BCPI-DNN*, *BPI-DNN*, *BCPI-RF* and *Marginal* attained a high performance. *Grouped Shapley* values, presented in additional experiments (section 7.6.2), showed a drop in the performance with the high-correlated settings. An extra simulation introducing more complexity with pair interactions of variables was conducted in additional experiments (sections 7.6.3 & 7.6.4).

The AUC score, type-I error, power and computation time for Experiment 7.3.2 are presented in Fig. 7.3.3. *BCPI-DNN* with *internal stacking* performed similarly to the same approach without stacking. Thus, both approaches showed comparable inferential behavior in identifying the significant groups. Nevertheless, in terms of computation time, the dimension reduction brought by stacking added significant benefits (around a factor of 2). In fact, in the *importance block* without stacking, all the variables of the remaining groups are used to predict those of the group of interest. Groups with high cardinality (of variables) are challenging in terms of memory resources and required computation, suggesting that *internal stacking* can help to reduce computational burden. The performance with groups of different cardinalities was conducted in the additional experiments (section 7.6.5).

Real-world empirical application of *BCPI-DNN* with *internal stacking* for age-prediction from brain imaging and socio-demographic information are summarized in Fig. 7.3.4. Results in **(Degree of Significance)** ranked the groups according to their corresponding level of significance. We choose a conservative significance level of $p = 0.001$ (Dashed line at $log_{10}(0.001) = 3$). Using the stacking approach, we scored the heterogeneous *brain* and *social* input variables regarding their predictive importance. As expected, we found that the brain groups - excluding *Brain DMRI MD* (see Table 7.7 for group description) - were highly important for age prediction. Interestingly, *Lifestyle* and *Education* were among the top predictive variables, conditional on the brain groups, suggesting the presence of complementary information. To challenge the plausibility of the selected groups, we investigated prediction performance after excluding non-significant groups. We used 10-fold cross validation with significance estimation and refitting the reduced model using the training set while scoring with the reduced model on the testing set. The reduced model did not perform visibly worse than the full model ($R^2 = 0.8, MAE = 2.9$), suggesting that our procedure effectively selects predictive groups. Of note the performance is in line with state-of-the art benchmarks on the UKBB based on convolutional neural networks ($MAE \sim$ 2-3 years, e.g., Roibu et al. [2023], Jonsson et al. [2019]). Consequently, results suggest that the proposed approach combined good prediction performance with effective identification of relevant groups of variables. Despite setting default behavior of the *internal stacking* approach to have one output neurone per group in the framework, a supplementary analysis considering multi-outputs per group is discussed in additional experiments (section 7.6.7).

7.5 . Discussion

In this work, we proposed *BCPI*, a novel and usable framework for computing single- and group-level variable importance. Our work provides statistical guarantees based on results from *Conditional Permutation Importance (CPI)*, whereas our implementation supports arbitrary regression and classification models consistent with the Scikit-learn API. We developed our approach by reproducing the known fact that standard *Permutation Importance (PI)*, represented by the *BPI-DNN* approach, lacks the ability to control type-I error [Williamson et al., 2021] with high correlated settings in Fig. 7.3.2, despite the high AUC score [Mi et al., 2021]. We extended these results, theoretically and empirically, to the group setting by proposing *BCPI-DNN*, which is built on top of an expressive DNN model as a base learner. This recipe led to high AUC scores while maintaining the control of type-I error across different correlation scenarios (Fig. 7.3.2).

Inspired by recent applications of model stacking for handling multiple

groups or input domains [Albu et al., 2023, Zhou et al., 2021, Engemann et al., 2020], we proposed *internal stacking* which implements stacking inside the DNN model, hence, avoiding separate optimization problems performed by common stacking pipelines. This was achieved by adding extra sub-linear layers to create a linear summary for each group of variables. Our benchmarks suggested that stacking maintained inferential performance of the full model while bringing time benefits (at least up to a factor of 2), especially for groups with high cardinality of variables (Fig. 7.3.3). Moreover, supplementary analyses of calibration of *BCPI-DNN* versus *BPI-DNN* (supplement, section 7.6.6) suggested that the p-values for *BCPI-DNN* showed a slightly conservative profile for *BCPI-DNN*. On the other hand, *BPI-DNN* showed poor calibration, once more underlining the relevance of conditional permutations.

Our empirical investigation of age prediction using heterogeneous inputs on the UKBB dataset suggests that the proposed framework facilitates constructing strong predictions models alongside trustworthy insights on the important predictive inputs. The prediction performance of our model was in line with state-of-the art benchmarks on the UKBB based on convolutional neural networks ($MAE \sim$ 2-3 years, e.g., [Roibu et al., 2023, Jonsson et al., 2019]) At the same time, the results provided a statistically grounded confirmation for the conclusions drawn in [Dadi et al., 2021] which were based on a less formal approach consistent with the *LOGI* approach.

Several limitations apply to our work. *BCPI-DNN* utilizes a DNN model as the base estimator for its high predictive accuracy. However, when the amount of training data is limited, the network can potentially memorize the training examples instead of learning generalizable patterns and a simpler base learner might be preferable, e.g. a Random Forest. By comparing the computation time for *BCPI-DNN* with the *internal stacking* approach between Fig. 7.3.3 and supplement Fig. 7.6.1 where the groups have a high cardinality (100) and a low cardinality (5) respectively, we can see that the use of *internal stacking* is preferable for high-cardinality situations. This is due to the extra training of the added sub-linear layers. Our work made use of pre-defined groups, which may not always be available. Instead, statistically defined groups could be used e.g. obtained from clustering algorithms. A possible issue might then be that the groups mix heterogeneous variables, which makes their interpretation challenging. On the flip side, reliance on pre-defined groups may lead to poor inference if the group structure does not track variable importance: if important variables are disseminated in all groups, the inference problem becomes much more challenging. This topic deserves careful investigation in the future. Moreover, here we only performed *internal stacking* by applying linear projection on the input data. It will be interesting to better understand the potential of non-linear projections.

Finally, additional possible future directions include studying the impact

of missing and low values on the accuracy, also across different group definitions. We hope that our results, resources and tools will facilitate the future study of the importance of groups of variables in prediction models.

7.6 . Additional Experiments

7.6.1 . Exp. 1 - Power & Computation time

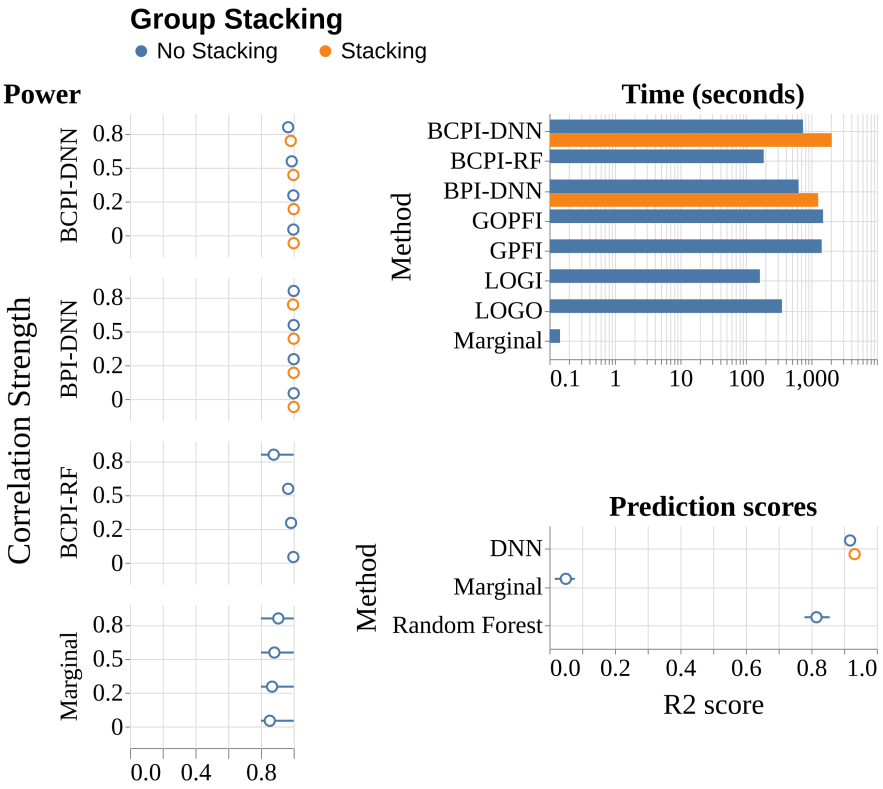

Figure 7.3.1-S1: **Benchmarking grouping methods**: *BCPI-DNN* is compared to baseline models and competing approaches for group variable importance providing p-values. **(Power)** indicates the mean proportion of informative variables identified. **(Time)** reports the computation time in seconds with $\log_{10}$ scale per core on 100 cores. **(Prediction scores)** presents the performance of the different base learners used in the group variable importance methods (*Marginal*: {Marginal effects}, *Random Forest*: {BCPI-RF, LOGI, LOGO, GPFI & GOPFI}, *DNN*: {BPI-DNN & BCPI-DNN}). Prediction tasks were simulated with n = 1000 and p = 50.

7.6.2 . Exp. 1 - AUC score for *Grouped Shapley* values

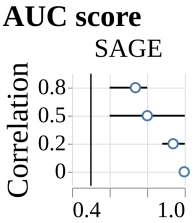

Figure 7.3.1-S2: ***Grouped Shapley* values**: Prediction tasks were simulated with n = 1000 and p = 50. Solid line: chance level.

The grouped version of SAGE (Global Importance with Shapley values [Covert et al., 2020]) was assessed with AUC scores (for detecting important variables) as it does not provide p-values. SAGE performed well in low-correlation settings (mean = 0.95) but the performance dropped in high-correlation settings (mean = 0.76).

7.6.3 . Exp. 1 - AUC score & Type-I error (Non linear case)

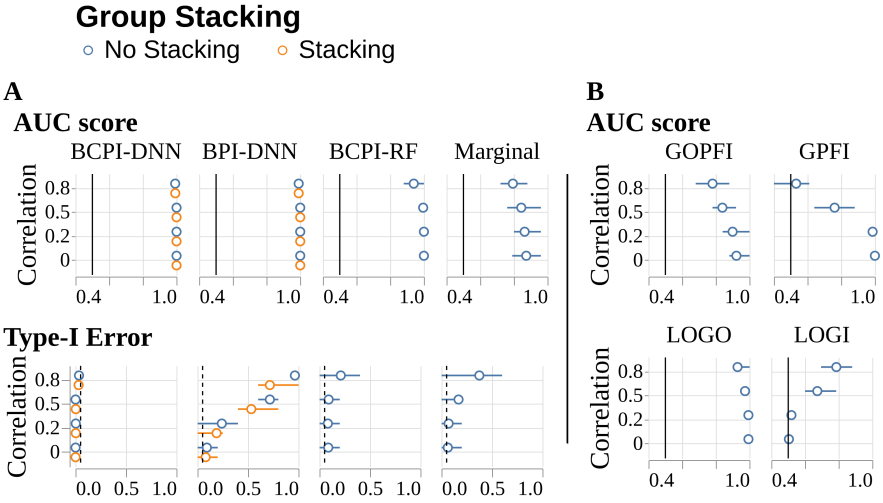

Figure 7.3.1-S3: **Benchmarking grouping methods**: *BCPI-DNN* is compared to baseline models and competing approaches for group variable importance. It encompasses two panels: **(A)** for the methods providing p-values used to check for AUC score and for statistical guarantees (Type-I error control), and **(B)** for the methods deprived of p-values, thus the importance scores are used to check for AUC score. Prediction tasks were simulated with n = 1000 and p = 50. Dashed line: targeted type-I error rate at 5%. Solid line: chance level.

To make the data-generating process more complex, we have added pair interactions to the regression simulation introduced in Fig. 7.3.2. The new out-

come is set to: $y_i = \mathbf{x_i}\boldsymbol{\beta^{main}} + \text{quad}(\mathbf{x_i}, \beta^{\mathbf{quad}}) + \sigma\epsilon_i,\ \forall i \in [\![n]\!]$ where the magnitude σ of the noise is set to $\frac{||\mathbf{X}\boldsymbol{\beta}^{\text{main}} + \text{quad}(\boldsymbol{X}, \boldsymbol{\beta^{quad}})||_2}{SNR\sqrt{n}}$ and $\text{quad}(\boldsymbol{x_i}, \boldsymbol{\beta^{quad}}) = \sum_{\substack{k,j=1 \\ k<j}}^{p_{signals}} \boldsymbol{\beta}^{\boldsymbol{quad}}_{k,j} x_i^k x_i^j$. The results show that *BCPI-DNN* outperforms all the alternatives methods presenting high AUC performance coupled with a control for type-I error under the predefined nominal rate. *BCPI-RF*, where the inference estimator is a Random Forest, showed an almost similar good performance with a little drop in high-correlated settings which can be explained by the drop in the predictive capacity following the plug of the Random Forest.

7.6.4 . Exp. 1 - Power & Computation time (Non linear case)

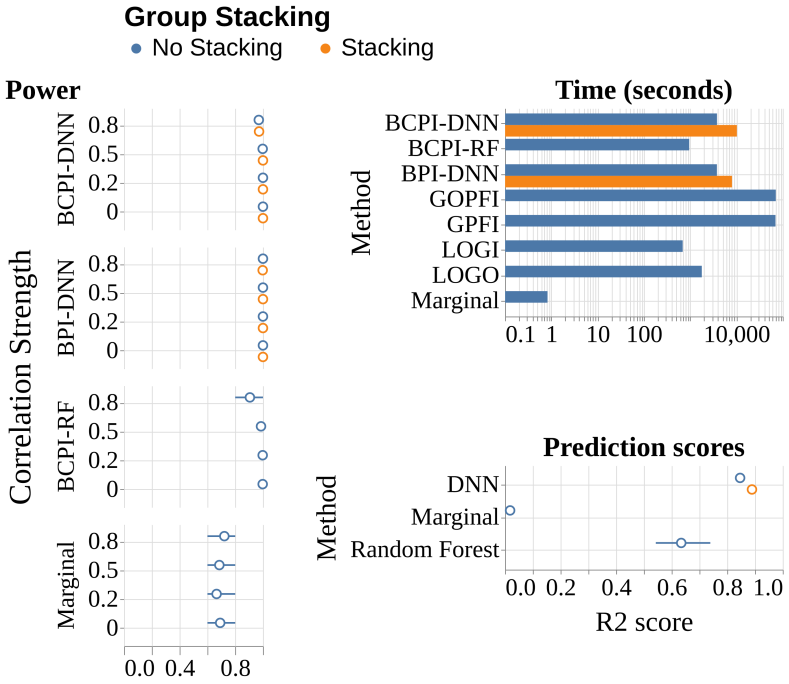

Figure 7.3.1-S4: **Benchmarking grouping methods**: *BCPI-DNN* is compared to baseline models and competing approaches for group variable importance providing p-values. **(Power)** indicates the mean proportion of informative variables identified. **(Time)** reports the computation time in seconds with $\log_{10}$ scale per core on 100 cores. **(Prediction scores)** presents the performance of the different base learners used in the group variable importance methods (*Marginal*: {Marginal effects}, *Random Forest*: {BCPI-RF, LOGI, LOGO, GPFI & GOPFI}, *DNN*: {BPI-DNN & BCPI-DNN}). Prediction tasks were simulated with n = 1000 and p = 50.

The results showed that *BCPI-DNN*, *BPI-DNN*, *BCPI-RF* and *Marginal* attained a high performance.

7.6.5 . Exp. 2 - Groups with different cardinalities

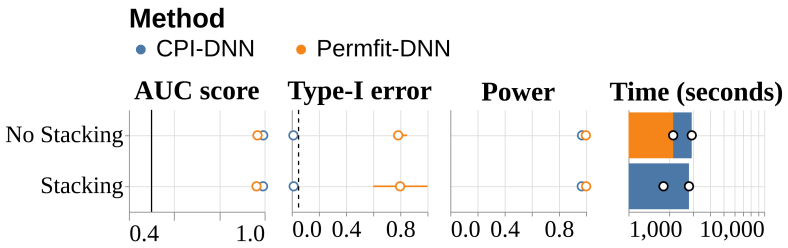

Figure 7.3.2-S1: **Groups of different cardinalities**: The performance of *BCPI-DNN* and *Permfit-DNN* at detecting important groups on simulated data with n = 1000 and p = 1000 with 10 blocks/groups, each group having a cardinality of 10 with or without the *stacking* approach. The **(AUC score)** evaluates the extent to which variables are ranked consistently with the ground truth. The **(Type-I error)** assesses the rate of low p-values (p-val < 0.05). **(Power)** provides information on the average proportion of detected informative variables (p-value < 0.05). The **(Time)** panel displays computation time in seconds with $\log_{10}$ scale per core on 100 cores. Dashed line: targeted type-I error rate. Solid line: chance level.

The results showed that *BCPI-DNN*'s capacity to achieve high AUC performance coupled with a control of Type-I error under the predefined nominal rate was maintained while providing groups of different cardinalities.

7.6.6 . Calibration of p-values between *BCPI-DNN* and *BPI-DNN*

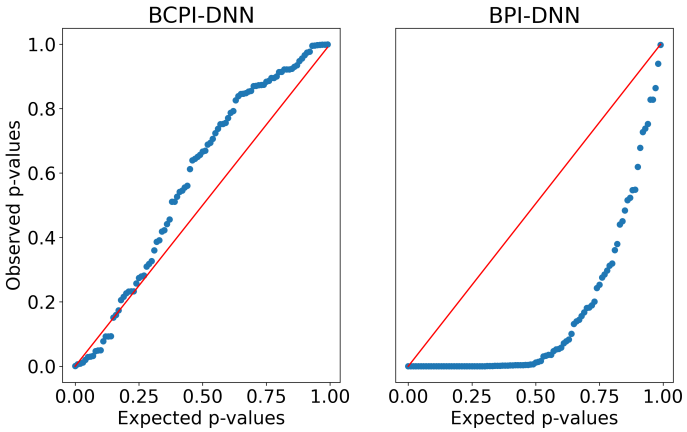

Figure 7-SE1: **P-values calibration**: The calibration of p-values ensuing from *BCPI-DNN* with the *conditional permutation* approach is compared to that of *BPI-DNN* with *standard permutation* approach. The p-value's distribution of one randomly selected non significant variable is compared to the uniform distribution. Prediction task was simulated with n = 1000 and p = 50.

7.6.7 . Impact of multi-output neurons on variable importance

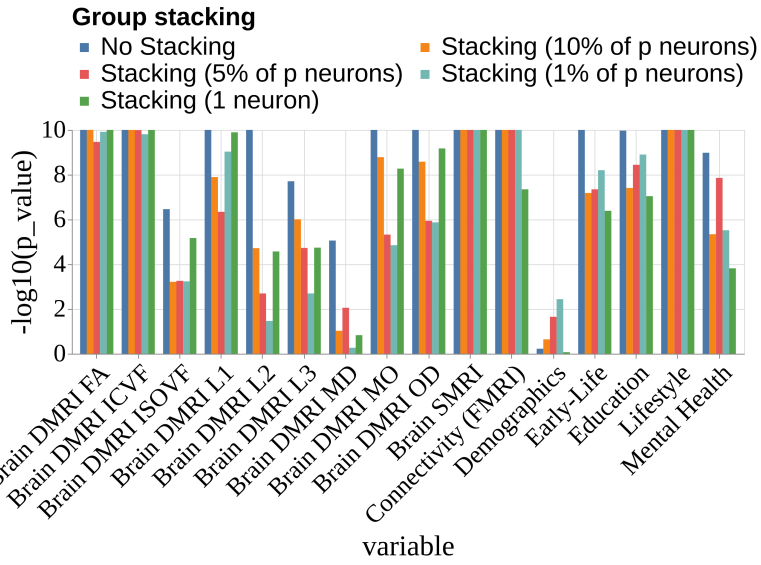

Figure 7-SE2: **Age prediction in UKBB with multi-output neurones**: Prediction of brain age from various socio-demographic and brain-imaging groups of phenotypes in a sample of n = 8357 volunteers from the UK BioBank. It plots the level of significance for the different groups in terms of -log10 of the derived p-values in terms of a changing number of output neurones for the *internal stacking* approach.

In Fig. 7-SE2, we compare the impact of having more output neurons following the linear projections with the suub-layers. The results indicated that the degree of significance of the different groups can change according to the level of information extracted per group, i.e. high vs low group cardinalities.

7.7 . Pre-defined groups in UK BioBank

Index	Name	# variables
1	Connectivity (FMRI)	1485
2	Brain DMRI FA	48
3	Brain DMRI ICVF	48
4	Brain DMRI ISOVF	48
5	Brain DMRI L1	48
6	Brain DMRI L2	48
7	Brain DMRI L3	48
8	Brain DMRI MD	48
9	Brain DMRI MO	48
10	Brain DMRI OD	48
11	Brain SMRI	157
12	Early-Life	8
13	Education	2
14	Lifestyle	45
15	Mental Health	25
16	Demographics	2

Table 7.1: **Knowledge-based groups in UK BioBank**: Imaging and socio-demographic formed groups within the data from UK Biobank with their corresponding cardinalities. *FMRI*: Functional Magnetic Resonance Imaging. Following Tae et al. [2018], Chen et al. [2016], *DMRI*: Diffusion Magnetic Resonance Imaging, *FA*: Fractional anisotropy (a measure of the degree of anisotropy of water diffusion in tissue), *ICVF*: Intra-Cellular Volume Fraction (a measure of the amount of space in tissue occupied by intracellular water), *ISOVF*: ISOtropic Volume Fraction (a measure of the amount of space in tissue occupied by freely diffusing water), *L1*: The largest eigenvalue of the diffusion tensor and indicates the rate of diffusion in the direction of the greatest diffusion, *L2*: An intermediate in size eigenvalue of the diffusion tensor and indicates the rate of diffusion in the direction perpendicular to the direction of the greatest diffusion, *L3*: The smallest eigenvalue of the diffusion tensor and indicates the rate of diffusion in the direction perpendicular to the first two directions, *MD*: Mean Diffusivity (a measure of the average rate of water diffusion in all directions), *MO*: Mode (a probabilistic tractography measure for crossing white matter fibers), *OD*: A measure of the angular difference between two sets of directions, *SMRI*: Structural Magnetic Resonance Imaging.

8 - Applications to Brain Imaging

Summary The tools developed in the previous chapters have the potential to enhance the application of machine learning in neuroscience. They provide a statistical framework that is rigorous and can be used to upgrade existing prediction models without imposing new types of architectures. This framework addresses the challenges in neuroscience and is applicable to both high-correlated and high-dimensional cases.

The absence of formal inference tools prevented promising ML applications from reaching insightful conclusions regarding the true parts of the decision-making process. Consequently, they were limited to the prediction capacity without statistically-based human comprehension. In this chapter, we revisit some results from the ML literature equipped with the built framework for variable importance with statistical guarantees, *BCPI*, within an attempt to provide clear answers to lingering questions related to the impact of the different predictors.

Throughout the next sections, we inserted both the standard permutation (*BPI-DNN*) and conditional permutation (*BCPI-DNN*) schemes in the prediction pipeline of existing real-world cases to monitor the impact of the corresponding predictors. This impact was measured by means of $-log_{10}$ of the group-wise p-values provided by the two methods respectively. In all the following experiments, to ensure that the selected groups are the statistically-based relevant ones, we performed a performance test following the removal of the detected non-important groups thresholding at 5% and 0.1% respectively.

8.1 . Exploring the influence of multimodal heterogeneous data on biomedical outcome prediction

8.1.1 . Challenge: inference of significant multimodal heterogeneous data

Several works made use of the brain phenotypes and socio-demographic data provided by the UKBB project with the aim of explaining the relationship between the health-based information and the individual traits [Gao et al., 2021, Kochunov et al., 2021, Mutz and Lewis, 2021]. These works demonstrate the pivotal role of accessible data in the generation ofhigh-quality predictions, thereby paving the way for the explanation of decision-making processes based on the various variables. With such big dataset ($n \sim 8300$ and $p \sim 2300$), iterating over all the variables in a perturbation-based process is costly, particularly for high-dimensional brain imaging data. Additionally, the high degree of correlation between brain imaging variables undermines the

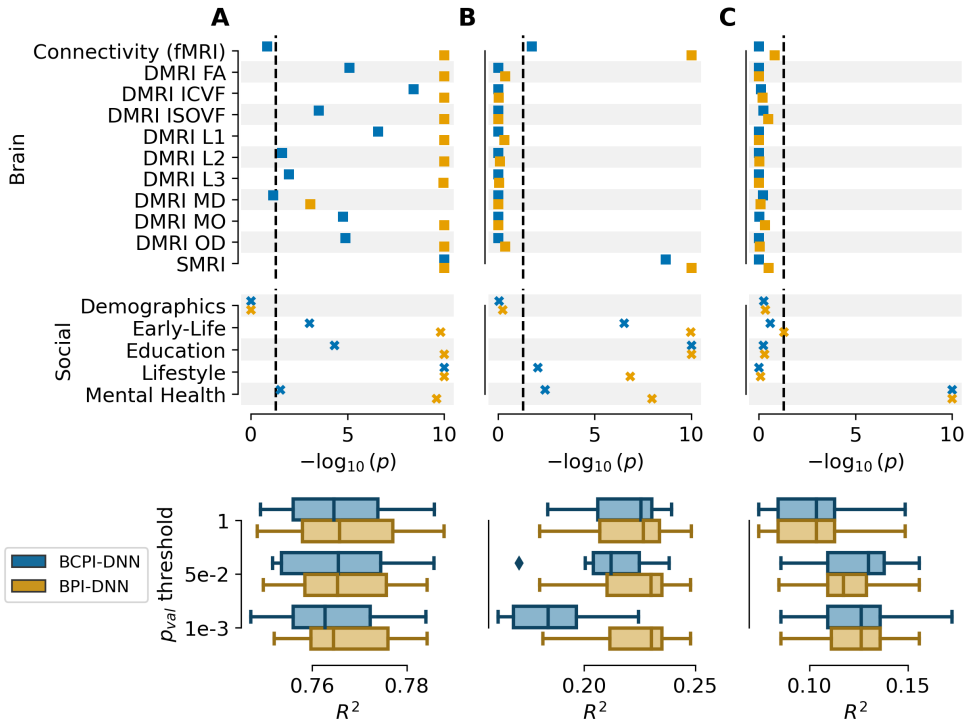

Figure 8.1.1: **Biomedical outcome prediction in UK Biobank:** Application of the standard permutation (*BPI-DNN*) and the conditional permutation (*BCPI-DNN*) to retrieve the impact of predefined brain and socio-demographic groups in UKBB for the prediction of **(A)** age, **(B)** fluid intelligence and **(C)** neuroticism respectively. (**Top Panel**) The degree of significance of the different groups by means of $-log_{10}$ of the p_{values}. (**Bottom Panel**) Predictive performance after removing the non-important groups (having p-value $>$ 0.05 or 0.001).

effectiveness of conditional permutation, as discussed in the previous chapter [XUE et al., 2010]. Thus, grouping is a proposed solution based on the pre-defined groups detailed in table 7.7 introduced in the work by Dadi et al. [2021]. A description of the UKBB dataset and its processing pipelines are provided in section 5.3.4 of chap. 5.

8.1.2 . Study & Results

We reanalyze the work by Dadi et al. [2021] on studying the impact of the mix of functional connectivity (fMRI) with Riemannian tangent-space embeddings (*Connectivity*), MRI diffusion-based data (*DMRI*) and socio-demographic data sources on the prediction of chronological age, fluid intelligence and neuroticism in the UK Biobank project. While presenting clear evidence of ML-based prediction, their work lacks any controlled statistical proof of the influence of the different inputs. Therefore, in Fig. 8.1.1, we conduct the same study while analyzing the marginal and conditional importance with statistical guarantees. We observed that both brain and socio-demographic factors had a statistically significant impact age prediction, whereas for fluid intelligence and neuroticism, the prediction was more reliant on socio-demographic factors, either partially or entirely. We should mention that the conditional scheme classified less groups as statistically relevant as compared to the permutation scheme, especially for the *connectivity* group considered as the group with the highest cardinality. This is due to the fact of the controlled selection of non-relevant predictors as relevant provided by the integration of the conditional scheme. As for the performance test, at the 5% level and across the different biomedical outcomes, we noted that the performance either remained constant or increased. At the 0.1% level, with age and neuroticism, we also watched a steady or an improved performance, while for intelligence the performance dropped significantly. One key explanation for this case is the level of fine-tuning of the integrated importance model, i.e. *Random Forest* model, making it highly-conservative with a negative effect on the final performance.

8.2 . Multimodal brain data: illuminating age prediction insights

8.2.1 . Challenge: inference of significant brain imaging modalities

Following the work by Engemann et al. [2020], predicting biomedical results through Magnetoencephalography (MEG) is pivotal in various applications such as monitoring neurodegenerative diseases, epilepsy surgery planning, or biomarker development, and is enhanced by supervised machine learning techniques. While the majority of the literature focused on the event-level outcomes, Sabbagh et al. [2020] concentrated on individual-based analysis predicting age under multimodal brain imaging modalities. A significant challenge in leveraging machine learning effectively within psychiatry and neurology, such as single-subject prediction from diverse neuroimaging modalities each with distinct data generation mechanisms, lies in the limited availability of extensive, high-quality datasets [Woo et al., 2017]. This challenge

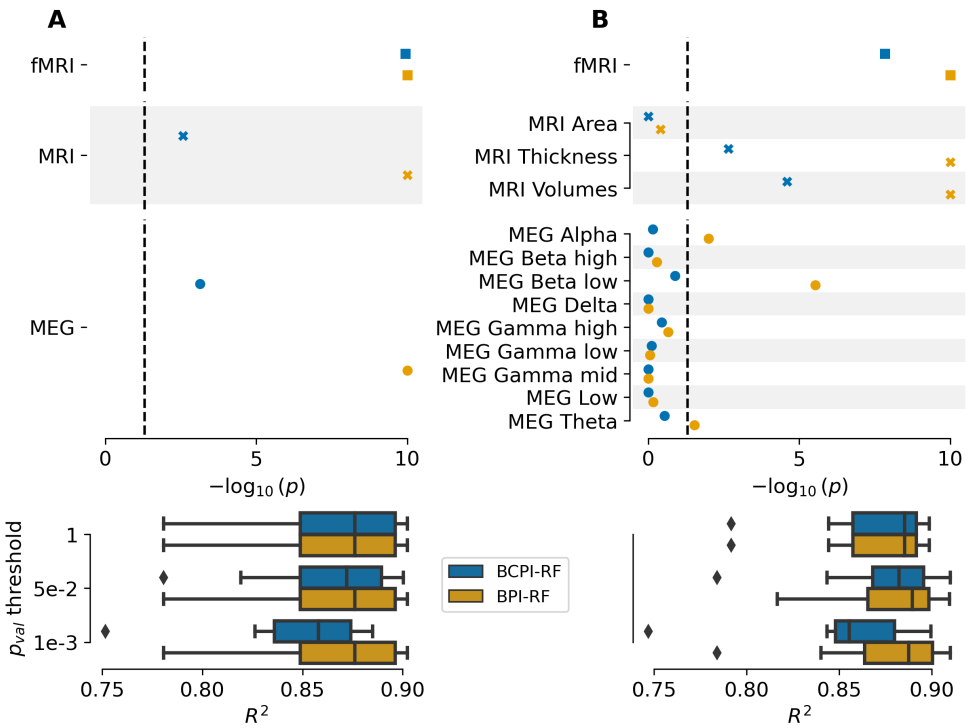

Figure 8.2.2: **Age prediction in Cam-CAN from Brain imaging modalities**: Application of the standard permutation (*BPI-DNN*) and the conditional permutation (*BCPI-DNN*) to retrieve the impact of brain imaging **(A)** modalities and **(B)** sub-modalities in Cam-CAN for the prediction of age. (Top Panel) The degree of significance of the different imaging modalities subgroups by means of $-log_{10}$ of the p_{values}. (Bottom Panel) Performance check after retrieving the non-important groups (having p-value > 0.05 or 0.001).

has been tackled recently with model-stacking techniques [Liem et al., 2017, Rahim et al., 2015]. In their study, Engemann et al. [2020] sought to identify the most informative electrophysiological markers of aging among fMRI, MRI and MEG. To this end, they imployed a traditional permutation-based importance approach, which has been demonstrated to provide non-controlled detection of relevant variables in high-correlated settings, thus achieving low-quality insights. Therefore, what remained was to integrate a statistically guaranteed importance indicator to the different brain modalities in the aforementioned pipeline. A description of the Cam-CAN dataset and its processing pipeline for MEG data are provided in section 5.5.3 of chap. 5. The following section provides a detailed overview of additional processing pipelines for MRI and fMRI data.

8.2.2 . Processing pipelines

MRI

A number of features listed below were extracted from MRI data using the established strategy outlined in the work by Liem et al. [2017]. This strategy relies on cortical surface reconstruction via FreeSurfer software.

Cortical thickness In a well-established association, cortical thinning is recognized as a hallmark of age-related brain atrophy [Thambisetty et al., 2010]. The cortical thickness is employed as a measure, defined as the distance between the white and pial mater surfaces. Cortical thickness data were extracted from FreeSurfer segmentation [Fischl, 2012], which utilizes a surface mesh with 5124 vertices in the standard fsaverage4 space. No further reduction in vertex count was implemented.

Cortical surface area Consistent with established findings of age-related decline [Lemaitre et al., 2012], the cortical surface area reduction was investigated. Estimates of cortical surface area at each vertex (vertex-wise) were computed by averaging the areas of faces surrounding each vertex on the white matter surface. This analysis was conducted using FreeSurfer segmentation software [Fischl, 2012] on a standardized surface mesh with 5124 vertices in fsaverage4 space. The mesh was employed without any further reduction in vertex count.

Subcortical volumes Building on the recognized association between aging and subcortical volume reduction [Murphy et al., 1992], the automated procedure using FreeSurfer generated 66 volumetric measures for each participant, without any further data reduction.

fMRI

Recent research has demonstrated that large-scale neuronal interactions between different brain networks undergo changes during healthy aging. To address the challenge of heterogeneity and dimensionality reduction, particularly in small- to medium-sized datasets like *Cam-CAN* with fewer than 1000 observations, fMRI-based predictive modeling has utilized functional atlases consisting of 50 to 1000 regions of interest (ROIs). These atlases serve as a foundational component [Dadi et al., 2019]. To estimate macroscopic functional connectivity, a departure was made from the 197-ROI BASC atlas [Bellec et al., 2010] used in the work by Liem et al. [2017]. Instead, a 256-ROI atlas with sparse and partially overlapping regions was adopted from Massive Online Dictionary Learning (MODL) [Mensch et al.]. Preliminary investigations indicated that both methods produced similar results, average with slightly reduced variance observed for the MODL atlas. Bivariate amplitude interactions were computed using Pearson correlations from the average time-series of each ROI. Subsequently, tangent space projection was employed to

vectorize the correlation matrices, resulting in $32,640$ connectivity values extracted from the lower triangle of each matrix. No additional reduction was performed.

8.2.3 . Study & Results

In Fig. 8.2.2, we studied the impact of different brain imaging modalities *MRI*, *fMRI* and *MEG* on the prediction of age for the participants in the cam-CAN study. The left panel studied the global impact of the different modalities by considering one group per modality, whereas the right panel further investigated the impact of the sub-groups composing each modality. We observed that the three brain modalities had a significant global impact on age prediction, and the significance score dropped when employing the conditional scheme. This result confirms the conclusion previously announced by Engemann et al. [2020] regarding the significance of integrating the three modalities for age prediction. At the sub-groups level, both permutations had an agreement with different scores for *fMRI* and *MRI*. However, for *MEG*, the standard permutation classified 3 sub-groups as relevant (Alpha, Beta low and Theta) while the conditional permutation didn't classify any sub-group as relevant. Although the permutation-based approach does validate the role of the MEG frequency bands that were previously identified in the original work, the conditional-based approach fails to do so. A plausible explanation is the small effects of the sub-groups that needs a more powerful importance model to capture it. Yet, we noticed that the performance following the removal of the non-relevant groups remained steady or slightly decreased. Further investigation is needed to pave the way for capturing the small effects which could introduce some improvements for the interpretation with the application of the conditional perturbation.

8.3 . Unlocking age prediction: insights from cortical brain regions

8.3.1 . Challenge: detection of predictive brain regions

Predicting age through brain scans involves examining the intricate folds of the cortex, the brain's convoluted outer layer responsible for functions like memory, focus, and higher cognition. Using MRI scans, scientists capture detailed images of the cortex, assessing its thickness, surface area, and volume across different brain regions in individuals spanning various age groups. Interestingly, these cortical attributes display discernible patterns as individuals age, reflecting both the brain's developmental trajectory and its natural aging process. Using machine learning techniques, researchers build computer models that analyze these patterns and learn to differentiate characteristics linked to various age groups, allowing them to predict an individual's

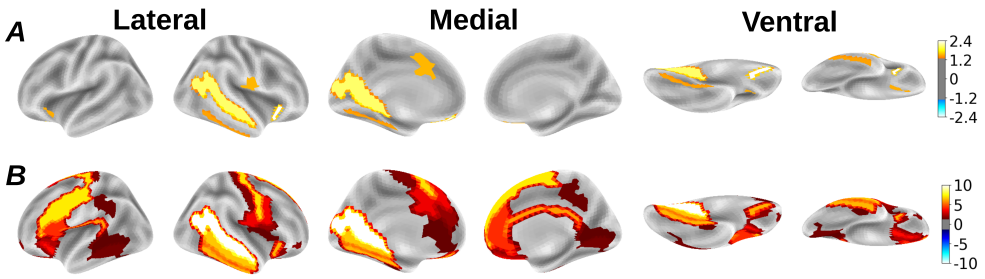

Figure 8.3.3: **Brain regions/parcels significantly contributing to age prediction**: Detection of predictive cortical brain regions for age prediction in Cam-CAN under different views. (**Surface**) shows the statistically validated relevant regions with their corresponding degree in terms of $-log_{10}$ of p_{values} under a threshold of 0.05. (**A**) and (**B**) stands respectively for the application of the conditional and standard permutations ($\mathbf{R^2}$ = 0.639, **MAE** = 8.606).

age with exceptional accuracy [Madan and Kensinger, 2018, Cole and Franke, 2017]. This method goes beyond just guessing age as it gives valuable information about how healthy brains age and provides a reference point for future comparisons [Teissier et al., 2020]. By comprehending the typical changes in a healthy cortex, scientists can differentiate between normal aging and the impacts of diseases that may accelerate or alter these patterns [Smith et al., 2015]. This research holds significant promise for the future, potentially revolutionizing neuroimaging studies by enhancing our understanding of how brain disorders influence cortical structure and aging [Gautherot et al., 2021]. Furthermore, this research isn't just about predicting age. It could also be used in clinics to help diagnose brain diseases [Franke and Gaser, 2019]. Ultimately, it sheds light on the amazing process of brain development and aging, and might even lead to new ways to keep our brains healthy throughout our lives [Koutsouleris et al., 2014]. In the following section, we revisit the work by Engemann et al. [2020] on identifying the impact of brain imaging modalities on age prediction without a statistically-based inference. Therefore, we focus on the influence of the cortical brain regions equipped with the developed framework to highlight the ability of detecting predictive parts.

8.3.2 . Study & Results

Brain imaging is a high-dimensional setting with hundreds of thousands of voxels [Bzdok et al., 2015] making it almost impossible to loop over each voxel to compute its impact on the prediction. As a result, in Fig. 8.3.3, we explored the influence of the different cortical regions on age prediction using the *cortical thickness* group in Cam-CAN. We used the atlas introduced in the work by Destrieux et al. [2010] gathering a sum of 75 Region of Interest (ROI). The results presented an agreement between both permutation schemes in classifying the *anterior and posterior subcentral sulci*, *inferior temporal gyrus*, *circular sulcus of the insula anterior*, *medial orbital sulcus* and *superior temporal sulcus* as relevant for the prediction of age while providing a decline in the degree of significance in comparison with the standard permutation. Nevertheless, the standard permutation scheme considered 20 more regions as relevant for this prediction which means the conditional perturbation highlighted less regions as relevant. The disagreement arose from the high correlation observed among the brain regions, as demonstrated in the work by Chevalier et al. [2021]. In this context, the conditional scheme has showcased its ability to more effectively regulate the true non-relevant predictors identified as relevant.

8.4 . Significant frequencies bands for age prediction

8.4.1 . Challenge: inference of the important frequencies from EEG models

Multiple studies have illustrated that EEG characteristics, such as rhythmic activity (e.g., delta, theta, alpha, beta, and gamma), vary with age [Al Zoubi et al., 2018, Cragg et al., 2011, Ashburner, 2007]. Sabbagh et al. [2020] discussed the effect of the EEG features' preprocessing for subject-level age prediction. The Source Power Comodulation (SPoC) is an algorithm that involves leveraging the information embedded in the outcome variable to direct the decomposition process, prioritizing source signals whose power correlates with the outcome [Dähne et al., 2014, Koles et al., 1990]. The authors introduced filterbank models derived from Riemannian geometry offering a viable substitute for MRI-based source localization, albeit with slightly inferior performance. Unlike individual-specific source estimates, where variations in head and sensor positions are explicitly accounted for, the Riemannian embedding assumes a constant linear field spread, which may not hold true across different recordings [Congedo et al., 2017]. These embeddings were used to measure the impact of EEG features on age prediction without retrieving the standalone effects of the different frequency bands. Al Zoubi et al. [2018] tended to search for the impact of each frequency band on the prediction of individual-based age. In their work, they used a thorough approach aimed at

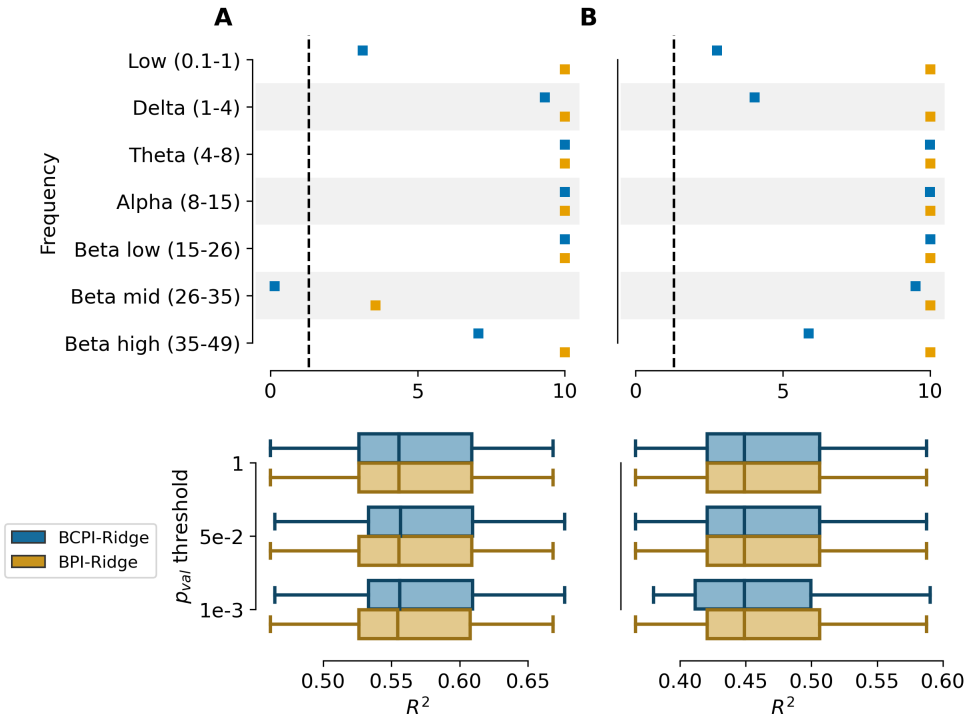

Figure 8.4.4: **Significant frequencies for age prediction under Riemannian/SPoC projectors in TUAB:** Application of the standard permutation (*BPI-DNN*) and the conditional permutation (*BCPI-DNN*) to retrieve the impact of different frequencies under **(A)** Riemannian and **(B)** SPoC projectors respectively in TU for the prediction of age. (**Left Panel**) The degree of significance of the different groups by means of $-log_{10}$ of the p_{values}. (**Right Panel**) Performance check after retrieving the non-important groups (having p-value $>$ 0.05 or 0.001).

examining all characteristics across every channel and frequency band without the reduction of features through averaging. The importance scores of the different predictors were fetched from the stack-ensemble model used in a nested cross validation manner without a statistical-base for the extracted insights. Using the filterbank provided embeddings along with statistically-based approach to measure the impact of the different frequency bands on subject-level age prediction remains an open question.

8.4.2 . Dataset description

The Temple University Hospital Abnormal EEG Corpus (TUAB/TUH-EEG) is a data collection initiative, including 14 years of clinical EEG data gathered at Temple University Hospital which offers a valuable resource for studying brain activity in a diverse population. This dataset primarily includes participants from Latin American and African American backgrounds [Obeid and Picone, 2016]. Each recording is subsequently labeled as either "normal" or "abnormal" by medical experts. Due to its clinical and social variety, TUAB is considered crucial for developing electrophysiological models that can be applied to diverse populations [Gemein et al., 2020, Sabbagh et al., 2020]. This study exclusively examined EEG recordings categorized as "normal" yielding a subset of 1385 individuals (775 females and 610 males). Ages ranged from newborn (minimum: 0 for females, 1 for males) to elderly (maximum: 95 for females, 90 for males) with an average age of 44.4 ± 16.5 years. The EEG data is provided in Volts with a standard deviation of 9.7 microvolts. The data processing procedures mirror the previous work by Engemann et al. [2022].

8.4.3 . Processing pipeline

EEG

EEG data were acquired using a variety of Nicolet EEG devices (Natus Medical Inc.) equipped with 24 to 36 channels. Channel placement followed the 10-5 system [Oostenveld and Praamstra, 2001] across all recordings. The initial sampling rate for the EEG data was 500 Hz. After applying a band-pass filter between 0.1 and 49 Hz, the data was resampled to a consistent rate of 200 Hz for further analysis. A common reference electrode was used during all recording sessions. We focused our analysis on a subset of 21 common channels. Due to inconsistencies in channel numbering across recordings, re-referencing with an average reference was necessary to ensure consistency across the entire EEG dataset. Additionally, variations in sampling frequencies necessitated resampling all data to a common rate of 200 Hz. In cases where multiple recordings were available for a participant, only the first recording was included for simplicity.

8.4.4 . Study & Results

We reanalyze the work in [Engemann et al., 2022, Sabbagh et al., 2020] on identifying the significant frequency bands from 0.1 Hz to 49 Hz for subject-level age prediction in the TUAB dataset. While this work provides insightful conclusions, it does not provide statistical-based inference about the impact of the frequency bands. Consequently, in Fig. 8.4.4, we conducted an examination of the marginal and conditional influence of the different frequency bands on chronological age. The features were extracted under the Riemannian and SPoC projectors respectively applying the filterbank models from

Sabbagh et al. [2020]'s work. With the Riemannian projector, the conditional permutation classified the *beta mid* frequency band as non-relevant as compared to the standard permutation. It offered also a drop in the degree of relevance for *low* and *beta high*. As for the SPoC projector, all the frequency bands were classified as relevant with a drop within the conditional permutation for *low*, *delta* and *beta high*. Thus, in both cases, the conditional permutation reduced the importance degree of *Low* and *beta high* frequency bands. Additionally, classifiying the *beta mid* band as non-relevant didn't affect the performance as shown in the bottom panel. These results are in agreement with the work by Bomatter et al. [2024] illustrated in Fig. 5. It is also noteworthy that the features obtained with the Riemannian embeddings improved the performance in terms of R^2 in comparison with SPoC.

8.5 . Significant frequencies for identifying the status of the eyes

8.5.1 . Challenge: inferring significant frequency contributions in EEG prediction models

The *Berger effect* is a well-established EEG phenomenon that refers to the increase in the alpha-band power over the occipital/parietal areas when individuals close their eyes compared to when they open their eyes [**?**Berger, 1969]. Barry et al. [2007] verified the utilization of the mean *alpha* level as an indicator of resting-state arousal during both eyes-closed and eyes-open states. Li [2010] uncovered the lack of impact of the *beta* bands in identifying the condition of the eyes. Michel et al. [2015] confirmed that *theta* waves indicate developmental shifts in the processing of eye gaze in infants. Assessing the influence of various frequency bands on individual-level eye state through the utilization of provided filterbank embeddings and statistical methodologies remains unresolved. The present study concerns a well-understood landmark, for which the 8-12 Hz band is expected to be of significance.

8.5.2 . Dataset description

The Leipzig Study for Mind-Body-Emotion Interactions (LEMON) focuses on a well-defined group of healthy young and elderly adults recruited from the general population, as described in the study by Babayan et al. [2019]. Similar to the Cam-CAN data, the LEMON research was conducted in a controlled research setting using high-quality equipment. Additionally, participants underwent extensive neurocognitive and behavioral assessments, providing valuable insights into their mental and behavioral functioning. The LEMON dataset incorporated resting-state Electroencephalography (EEG) from 227 healthy individuals consisting of 82 females (mean age = 42) and 145 males (mean age = 36), indicating a noticeable disparity in gender representation.

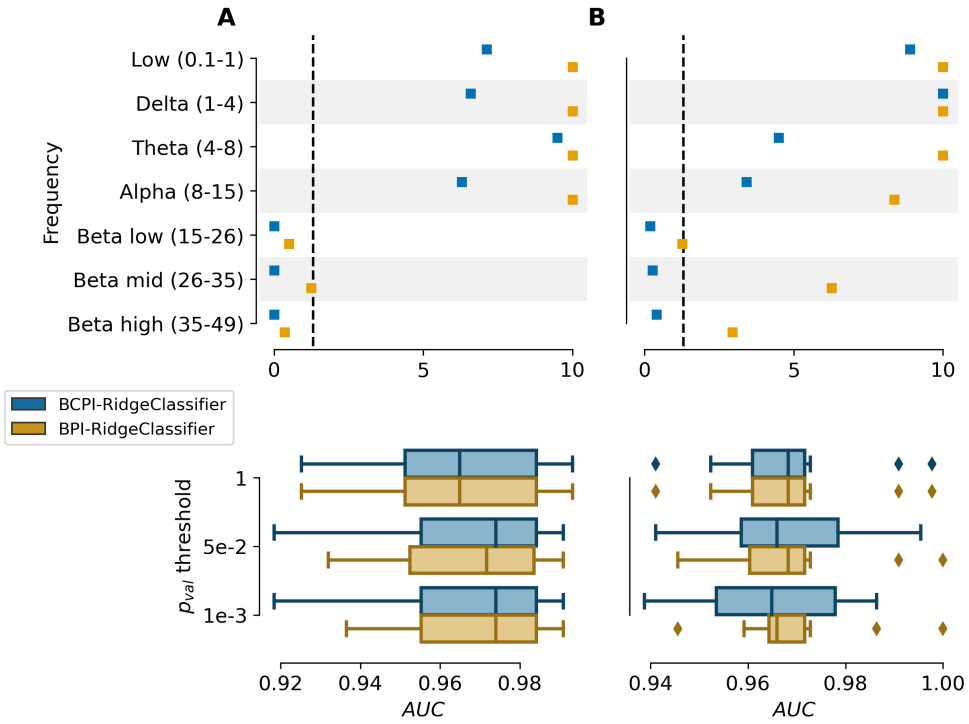

Figure 8.5.5: **Significant frequencies for eyes' status under Riemannian/SPoC projectors in LEMON:** Application of the standard permutation (*BPI-DNN*) and the conditional permutation (*BCPI-DNN*) to retrieve the impact of different frequencies under **(A)** Riemannian and **(B)** SPoC projectors respectively in LEMON for the prediction of the status of the eyes. (**Left Panel**) The degree of significance of the different groups by means of $-log_{10}$ of the p_{values}. (**Right Panel**) Performance check after retrieving the non-important groups (having p-value $>$ 0.05 or 0.001).

The age range spanned from 20 to 77 years, with an average age of 38.9 ± 20.3 years. An important aspect of the LEMON dataset is its division of participants into two specific age ranges: 20-35 years and 55-77 years. This structure makes the overall average age a less accurate representation of the true age distribution. Furthermore, the publicly available version of the dataset offers age data only in 5-year intervals as a measure to safeguard participant confidentiality, although their impact on the modeling results proved to be minimal [Engemann et al., 2022]. Finally, the EEG data is provided in Volts with a standard deviation of 9.1 microvolts.

8.5.3 . Processing pipeline

EEG

EEG data acquisition employed a 62-channel active electrode system (ActiCAP) with an initial band-pass filter set between 0.015 Hz and 1 kHz. To further refine the signal, an additional band-pass filter between 0.1 Hz and 49 Hz was implemented. Consistent with the previous dataset, channel placement adhered to the 10-5 system [Oostenveld and Praamstra, 2001]. The data were initially sampled at a high rate of 2500 Hz. Subsequently, to improve manageability for further analysis, the data were decimated by a factor of 5, yielding a final sampling frequency of 500 Hz. It is noteworthy that resting-state recordings within this dataset included samples from two distinct conditions: eyes closed and eyes open. The data processing pipeline explicitly considered these separate conditions. To maximize data utilization and potentially identify a wider variety of distinguishable EEG sources, the data from both conditions were pooled prior to feature extraction.

8.5.4 . Study & Results

In Fig. 8.5.5, we studied the corresponding frequency bands responsible for determining the status of the eyes (open/close) in a classification-based problem. Again, we deploy the filterbank models implemented in the work by Sabbagh et al. [2020] to attain the projected embeddings of the EEG features. Through the Riemannian projector, both permutation schemes reached a good agreement to classify the different frequency bands into relevant and non-relevant categories. All the *beta* bands were considered as non-relevant with a slight difference in the degree of significance. Additionally, the conditional permutation pointed a drop in the degree for *low*, *delta* and *alpha*. Regarding the SPoC projector, this agreement didn't remain solid with the standard perturbation highlighting the relevance of the *beta* bands. A bigger drop in the degree was watched for both *theta* and *alpha* bands, as the conditional alternative is characterized by controlling for type-I error. Consequently, with both projectors, *alpha* is a relevant frequency band for the condition of the eyes. These results are consistent with the previous work in the literature. The removal of the non-relevant frequency bands with both projectors had no effect on the performance measured in terms of *AUC* score. The EEG features extracted with both projectors achieved a high *AUC* score of 0.97.

8.6 . Conclusion

The application of machine learning (ML) techniques with complex, high-capacity models is becoming increasingly prevalent in the field of neuroscience. However, there is a lack of understanding regarding the effect of different predictors on the prediction of biomedical outcomes with statistical guarantees. While the standard permutation scheme is straightforward to implement, it can lead to the misinterpretation of non-relevant predictors as relevant. Consequently, in this study, we employed the block-based conditional scheme (BCPI) to regulate the false positive rate (FPR) with four distinct biomedical datasets: *UK Biobank*, *Cam-CAN*, *TUAB*, and *LEMON*. With *UK Biobank*, the method validated statiscally the utility of an entire or partial mixture of brain and socio-demographic data for the prediction of age, fluid intelligence and neuroticism, as discussed in the previous work, while offering a statiscally validated reduction in the number of variables. Regarding the *Cam-CAN* dataset comprising multimodal brain data, the fMRI, MRI, and MEG frequency bands had a global impact on age prediction, which confirms the impact of combining the three brain imaging modalities. However, there was a notable absence of the role of the MEG modality's frequency bands captured by the conditional approach in comparison to the marginal one. This is due to the fact that the signal is distributed, and measuring separate variables may result in detecting only small effects. With regard to the *TUAB* and *LEMON* datasets, where the frequency analysis is of interest for both age prediction and the state of the eyes, the Riemannian projected embeddings of the EEG features were found to be more suitable for the predictions, indicating that the beta bands were not involved.

9 - Conclusion

With machine learning (ML) becoming increasingly prevalent in scientific domains such as neuroscience, and the integration of high-capacity models capable of detecting non-linear patterns in data and generating high-quality predictions, there is a growing need to gain a deeper understanding of its decision-making process from a human perspective in order to draw meaningful conclusions. However, the current status of XAI literature lacks the requested tools with a highly sensitive statistically rigorous method controlling for flaws in the detection of relevant variables. Therefore, in this thesis, we have developped a framework for variable importance with statistical guarantees. It is applicable for both single and group levels within large, high-dimensional datasets. The scenario we have addressed involves a combination of high-dimensionality and high-correlation cases, which are two major challenges in neuroscience.

First, in chap. 5, we discussed the theoretical limitations of the standard permutation importance (*Permfit*) approach, which does not control the type-I error rate with correlated data. We then presented a new algorithm for conditional permutation importance (*CPI*), which assesses the impact on the target loss function when a variable is permuted. However, the permutation is applied specifically to the residuals of the regression of the variable of interest on the others, resulting in a conditional approach. A series of simulations have demonstrated that this procedure is accurately controlling type-I error while achieving a high AUC score, which is based on a DNN learner. Our application on the big heterogeneous biomedical dataset *UK Biobank* has highlighted the efficiency of the method. It detects fewer significant variables, which is compatible with the theoretical results. While CPI overcomes the limitation of standard Permfit in high-correlated settings, a new limitation arises from the definition of the conditional inference. With two extremely correlated and significant variables, their conditional importance is cancelled out, leading to their identification as non-significant.

Next, in chap. 7, we extended the aforementioned approach to introduce an improved framework, block-based conditional permutation importance (*BCPI*), which utilises a prior grouping of the variables. This grouping may be either data-driven or domain knowledge-driven. Given the time-consuming nature of dealing with high-dimensional cases, we introduced a new stacking approach that expanded the architecture of the DNN learner to integrate sub-linear layers. This linearly projecting the variables of each group to a low-dimensional trainable space. Through a series of simulations, we demonstrated that BCPI maintains control of type-I error with a high AUC score across different correlation scenarios. Furthermore, the novel stacking ap-

proach enabled the preservation of the achieved performance while offering an advantageous improvement in time-consumption. Our empirical study on age prediction using diverse inputs from the UK Biobank dataset indicates that the proposed framework has enabled the development of robust prediction models with the identification of the significant predictive inputs. The reliance on pre-defined groups may result in suboptimal inference if the group structure does not accurately reflect variable importance. If important variables are distributed across all groups, addressing the inference challenge becomes significantly more difficult, which merits further investigation in future research. Additionally, our current approach involves only internal stacking, achieved through linear projection on the input data. Exploring the potential of non-linear projections would be of great interest. Furthermore, a potential future direction involves examining how missing and low values affect accuracy, as well as exploring various group definitions.

Finally, having acquired the requisite tools for both high-correlation and high-dimensional scenarios, we return to the primary objective of the community: the extraction meaningful interpretations from real-world biomedical datasets composed of diverse sources of data. In chap. 8, we apply the developed framework for variable importance, equipped with statistical guarantees BCPI on four big multimodal biomedical cohorts: *UK Biobank*, *Cam-CAN*, *TUAB* and *LEMON*. The framework is suitable for drawing inferences on correlated input variables and even on entire groups of variables simultaneously. With the UK Biobank dataset, the method demonstrated the efficacy of integrating a combination of brain and socio-demographic data in order to predict age, fluid intelligence, and neuroticism, while also statistically reducing the number of variables. In the Cam-CAN dataset, which includes multimodal brain data, such as fMRI, MRI, and MEG frequency bands, age prediction was significantly impacted. However, the frequency bands within the MEG modality appeared to be of lesser importance, suggesting a need for further investigation to capture small effects distributed among frequency bands. Additionally, in the TUAB and LEMON datasets, where frequency analysis is crucial for age prediction and eye state assessment, EEG features obtained by Riemannian emebddings features outperformed other methods. This is also particularly highlighted the limited involvement of beta bands.

This thesis makes a compelling case for the community of researchers to utilize this tool to address unresolved conclusions for ML applications, thereby enhancing the potential for new discoveries.

References

Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform*, 8:14, 2014. ISSN 1662-5196. doi: 10.3389/fninf.2014.00014.

Charu C. Aggarwal and Jiawei Han, editors. *Frequent Pattern Mining*. Springer, 2014. ISBN 978-3-319-07820-5. doi: 10.1007/978-3-319-07821-2.

Serkan Akogul. A Novel Approach to Increase the Efficiency of Filter-Based Feature Selection Methods in High-Dimensional Datasets With Strong Correlation Structure. *IEEE Access*, 11:115025–115032, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3325331.

Obada Al Zoubi, Chung Ki Wong, Rayus T. Kuplicki, Hung-wen Yeh, Ahmad Mayeli, Hazem Refai, Martin Paulus, and Jerzy Bodurka. Predicting Age From Brain EEG Signals—A Machine Learning Approach. *Front. Aging Neurosci.*, 10, July 2018. ISSN 1663-4365. doi: 10.3389/fnagi.2018.00184.

Mark Alber, Adrian Buganza Tepole, William R. Cannon, Suvranu De, Salvador Dura-Bernal, Krishna Garikipati, George Karniadakis, William W. Lytton, Paris Perdikaris, Linda Petzold, and Ellen Kuhl. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digit. Med.*, 2(1): 1–11, November 2019. ISSN 2398-6352. doi: 10.1038/s41746-019-0193-y.

Alexandra-Ioana Albu, Maria-Iuliana Bocicor, and Gabriela Czibula. MM-StackEns: A new deep multimodal stacked generalization approach for protein–protein interaction prediction. *Computers in Biology and Medicine*, 153:106526, February 2023. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2022.106526.

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99:101805, November 2023. ISSN 1566-2535. doi: 10.1016/j.inffus.2023.101805.

Marlene Sophie Altenmüller, Leonie Lucia Lange, and Mario Gollwitzer. When research is me-search: How researchers' motivation to pursue a topic affects laypeople's trust in science. *PLoS One*, 16(7):e0253911, July 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0253911.

André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, May 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq134.

Anup Amatya and Hakan Demirtas. OrdNor: An R Package for Concurrent Generation of Correlated Ordinal and Normal Data. *Journal of Statistical Software*, 68:1–14, November 2015. ISSN 1548-7660. doi: 10.18637/jss.v068.c02.

Melis Anatürk, Tobias Kaufmann, James H. Cole, Sana Suri, Ludovica Griffanti, Enikő Zsoldos, Nicola Filippini, Archana Singh-Manoux, Mika Kivimäki, Lars T. Westlye, Klaus P. Ebmeier, and Ann-Marie G. de Lange. Prediction of brain age and cognitive age: Quantifying brain and cognitive maintenance in aging. *Hum Brain Mapp*, 42(6):1626–1640, April 2021. ISSN 1097-0193. doi: 10.1002/hbm.25316.

Jesper L.R. Andersson and Stamatios N. Sotiropoulos. Non-parametric representation and prediction of single- and multi-shell diffusion-weighted MRI data using Gaussian processes. *Neuroimage*, 122:166–176, November 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.07.067.

Daniel W. Apley and Jingyu Zhu. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models, August 2019.

John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, October 2007. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2007.07.007.

Quay Au, Julia Herbinger, Clemens Stachl, Bernd Bischl, and Giuseppe Casalicchio. Grouped Feature Importance and Combined Features Effect Plot, April 2021.

Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D. Reinelt, Andrea M. F. Reiter, Josefin Röbbig, H. Lina Schaare, Marie Uhlig, Alfred Anwander, Pierre-Louis Bazin, Annette Horstmann, Leonie Lampe, Vadim V. Nikulin, Hadas Okon-Singer, Sven Preusser, André Pampel, Christiane S. Rohr, Julia Sacher, Angelika Thöne-Otto, Sabrina Trapp, Till Nierhaus, Denise Altmann, Katrin Arelin, Maria Blöchl, Edith Bongartz, Patric Breig, Elena Cesnaite, Sufang Chen, Roberto Cozatl, Saskia Czerwonatis, Gabriele Dambrauskaite, Maria Dreyer, Jessica Enders, Melina Engelhardt, Marie Michele Fischer, Norman Forschack, Johannes Golchert, Laura Golz, C. Alexandrina Guran, Susanna Hedrich, Nicole Hentschel, Daria I. Hoffmann, Julia M. Huntenburg, Rebecca Jost, Anna Kosatschek, Stella Kunzendorf, Hannah Lammers, Mark E. Lauckner, Keyvan Mahjoory, Ahmad S. Kanaan, Natacha Mendes, Ramona

Menger, Enzo Morino, Karina Näthe, Jennifer Neubauer, Handan Noyan, Sabine Oligschläger, Patricia Panczyszyn-Trzewik, Dorothee Poehlchen, Nadine Putzke, Sabrina Roski, Marie-Catherine Schaller, Anja Schieferbein, Benito Schlaak, Robert Schmidt, Krzysztof J. Gorgolewski, Hanna Maria Schmidt, Anne Schrimpf, Sylvia Stasch, Maria Voss, Annett Wiedemann, Daniel S. Margulies, Michael Gaebler, and Arno Villringer. A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Sci Data*, 6(1):180308, February 2019. ISSN 2052-4463. doi: 10.1038/sdata.2018.308.

Fakhirah Badrulhisham, Esther Pogatzki-Zahn, Daniel Segelcke, Tamas Spisak, and Jan Vollert. Machine learning and artificial intelligence in neuroscience: A primer for researchers. *Brain, Behavior, and Immunity*, 115:470–479, January 2024. ISSN 0889-1591. doi: 10.1016/j.bbi.2023.11.005.

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5), October 2015. ISSN 0090-5364. doi: 10.1214/15-AOS1337.

Robert J. Barry, Adam R. Clarke, Stuart J. Johnstone, Christopher A. Magee, and Jacqueline A. Rushby. EEG differences between eyes-closed and eyes-open resting conditions. *Clinical Neurophysiology*, 118(12):2765–2773, December 2007. ISSN 1388-2457. doi: 10.1016/j.clinph.2007.07.028.

Danielle S. Bassett and Michael S. Gazzaniga. Understanding complexity in the human brain. *Trends Cogn Sci*, 15(5):200–209, May 2011. ISSN 1364-6613. doi: 10.1016/j.tics.2011.03.006.

Mr. Bayes and Mr. Price. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions (1683-1775)*, 53:370–418, 1763. ISSN 0260-7085.

Christian F. Beckmann and Stephen M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans Med Imaging*, 23(2):137–152, February 2004. ISSN 0278-0062. doi: 10.1109/TMI.2003.822821.

Pierre Bellec, Pedro Rosa-Neto, Oliver C. Lyttelton, Habib Benali, and Alan C. Evans. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *Neuroimage*, 51(3):1126–1139, July 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.02.082.

H. Berger. On the electroencephalogram of man. *Electroencephalogr Clin Neurophysiol*, pages Suppl 28:37+, 1969. ISSN 0013-4694.

Przemyslaw Biecek. DALEX: Explainers for complex predictive models. https://arxiv.org/abs/1806.08915v2, June 2018.

B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn Reson Med*, 34(4):537–541, October 1995. ISSN 0740-3194. doi: 10.1002/mrm.1910340409.

Kristin Blesch, David S. Watson, and Marvin N. Wright. Conditional Feature Importance for Mixed Data. *AStA Adv Stat Anal*, April 2023. ISSN 1863-8171, 1863-818X. doi: 10.1007/s10182-023-00477-9.

Philipp Bomatter, Joseph Paillard, Pilar Garces, Jörg Hipp, and Denis Engemann. Machine learning of brain-specific biomarkers from EEG, January 2024.

Joachim Böttger, Alexander Schäfer, Gabriele Lohmann, Arno Villringer, and Daniel S. Margulies. Three-dimensional mean-shift edge bundling for the visualization of functional connectivity in the brain. *IEEE Trans Vis Comput Graph*, 20(3):471–480, March 2014. ISSN 1941-0506. doi: 10.1109/tvcg.2013.114.

Matthew Bracher-Smith, Elliott Rees, Georgina Menzies, James T. R. Walters, Michael C. O'Donovan, Michael J. Owen, George Kirov, and Valentina Escott-Price. Machine learning for prediction of schizophrenia using genetic and demographic factors in the UK biobank. *Schizophrenia Research*, 246:156–164, August 2022. ISSN 0920-9964. doi: 10.1016/j.schres.2022.06.006.

Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997. ISSN 0031-3203. doi: 10.1016/S0031-3203(96)00142-2.

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

Peter Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4), September 2013. ISSN 1350-7265. doi: 10.3150/12-BEJSP11.

Ninon Burgos. Neuroimaging in Machine Learning for Brain Disorders. In Olivier Colliot, editor, *Machine Learning for Brain Disorders*, pages 253–284. Springer US, New York, NY, 2023. ISBN 978-1-07-163195-9. doi: 10.1007/978-1-0716-3195-9_8.

Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. December 2020.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0579-z.

Danilo Bzdok, Michael Eickenberg, Olivier Grisel, Bertrand Thirion, and Gael Varoquaux. Semi-Supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection. *arXiv:1610.02351 [math, stat]*, December 2017.

Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl. Visualizing the Feature Importance for Black Box Models. In Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 655–670, Cham, 2019. Springer International Publishing. ISBN 978-3-030-10925-7. doi: 10.1007/978-3-030-10925-7_40.

Diego Castillo-Barnes, Javier Ramírez, Fermín Segovia, Francisco J. Martínez-Murcia, Diego Salas-Gonzalez, and Juan M. Górriz. Robust Ensemble Classification Methodology for I123-Ioflupane SPECT Images and Multiple Heterogeneous Biomarkers in the Diagnosis of Parkinson's Disease. *Frontiers in Neuroinformatics*, 12:53, 2018. ISSN 1662-5196. doi: 10.3389/fninf.2018.00053.

Debrup Chakraborty and Nikhil R. Pal. Selecting Useful Groups of Features in a Connectionist Framework. *IEEE Transactions on Neural Networks*, 19(3): 381–396, March 2008. ISSN 1941-0093. doi: 10.1109/TNN.2007.910730.

Ahmad Chamma, Denis A. Engemann, and Bertrand Thirion. Statistically Valid Variable Importance Assessment through Conditional Permutations, October 2023.

Nilanjan Chatterjee and Raymond J. Carroll. Semiparametric Maximum Likelihood Estimation Exploiting Gene-Environment Independence in Case-Control Studies. *Biometrika*, 92(2):399–418, 2005. ISSN 0006-3444.

Geng Chen, Pei Zhang, Ke Li, Chong-Yaw Wee, Yafeng Wu, Dinggang Shen, and Pew-Thian Yap. Improving Estimation of Fiber Orientations in Diffusion

MRI Using Inter-Subject Information Sharing. *Sci Rep*, 6(1):37847, November 2016. ISSN 2045-2322. doi: 10.1038/srep37847.

Yunsong Chen, Xiaogang Wu, Anning Hu, Guangye He, and Guodong Ju. Social prediction: A new research paradigm based on machine learning. *J. Chin. Sociol.*, 8(1):15, September 2021. ISSN 2198-2635. doi: 10.1186/s40711-021-00152-z.

Jérôme-Alexis Chevalier, Tuan-Binh Nguyen, Joseph Salmon, Gaël Varoquaux, and Bertrand Thirion. Decoding with confidence: Statistical control on decoder maps. *NeuroImage*, 234:117921, July 2021. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2021.117921.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1), March 2010. ISSN 1932-6157. doi: 10.1214/09-AOAS285.

D. Cohen. Magnetoencephalography: Detection of the brain's electrical activity with a superconducting magnetometer. *Science*, 175(4022):664–666, February 1972. ISSN 0036-8075. doi: 10.1126/science.175.4022.664.

James H. Cole and Katja Franke. Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. *Trends Neurosci*, 40(12):681–690, December 2017. ISSN 1878-108X. doi: 10.1016/j.tins.2017.10.001.

Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4:1–20, March 2017. doi: 10.1080/2326263X.2017.1297192.

Andrei-Emil Constantinescu, Ruth E. Mitchell, Jie Zheng, Caroline J. Bull, Nicholas J. Timpson, Borko Amulic, Emma E. Vincent, and David A. Hughes. A framework for research into continental ancestry groups of the UK Biobank. *Human Genomics*, 16(1):3, January 2022. ISSN 1479-7364. doi: 10.1186/s40246-022-00380-5.

Andrea Coravos, Sean Khozin, and Kenneth D. Mandl. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *npj Digit. Med.*, 2(1):1–5, March 2019. ISSN 2398-6352. doi: 10.1038/s41746-019-0090-4.

Ian Covert, Scott Lundberg, and Su-In Lee. Understanding Global Feature Contributions With Additive Importance Measures, October 2020.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by Removing: A Unified Framework for Model Explanation, May 2022.

Lucy Cragg, Natasa Kovacevic, Anthony Randal McIntosh, Catherine Poulsen, Kristina Martinu, Gabriel Leonard, and Tomáš Paus. Maturation of EEG power spectra in early adolescence: A longitudinal study. *Dev Sci*, 14(5): 935–943, September 2011. ISSN 1467-7687. doi: 10.1111/j.1467-7687.2010.01031.x.

Robert A. Cribbie. Evaluating the importance of individual parameters in structural equation modeling: The need for type I error control. *Personality and Individual Differences*, 29(3):567–577, September 2000. ISSN 0191-8869. doi: 10.1016/S0191-8869(99)00219-6.

Kamalaker Dadi, Mehdi Rahim, Alexandre Abraham, Darya Chyzhyk, Michael Milham, Bertrand Thirion, Gaël Varoquaux, and Alzheimer's Disease Neuroimaging Initiative. Benchmarking functional connectome-based predictive models for resting-state fMRI. *Neuroimage*, 192:115–134, May 2019. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2019.02.062.

Kamalaker Dadi, Gaël Varoquaux, Josselin Houenou, Danilo Bzdok, Bertrand Thirion, and Denis Engemann. Population modeling with machine learning can enhance measures of mental health. *GigaScience*, 10(10):giab071, October 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab071.

Alessandro Daducci, Erick J. Canales-Rodríguez, Hui Zhang, Tim B. Dyrby, Daniel C. Alexander, and Jean-Philippe Thiran. Accelerated Microstructure Imaging via Convex Optimization (AMICO) from diffusion MRI data. *Neuroimage*, 105:32–44, January 2015. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2014.10.026.

Sven Dähne, Frank C. Meinecke, Stefan Haufe, Johannes Höhne, Michael Tangermann, Klaus-Robert Müller, and Vadim V. Nikulin. SPoC: A novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. *Neuroimage*, 86:111–122, February 2014. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2013.07.079.

Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Madhu Babu, and Mohamed Jawed Ahsan. Machine Learning in Drug Discovery: A Review. *Artif Intell Rev*, 55(3):1947–1999, 2022. ISSN 0269-2821. doi: 10.1007/s10462-021-10058-4.

Dries Debeer and Carolin Strobl. Conditional permutation importance revisited. *BMC Bioinformatics*, 21(1):307, July 2020. ISSN 1471-2105. doi: 10.1186/s12859-020-03622-2.

Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul

Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, July 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.01.021.

Marie-Ève Desjardins, Julie Carrier, Jean-Marc Lina, Maxime Fortin, Nadia Gosselin, Jacques Montplaisir, and Antonio Zadra. EEG Functional Connectivity Prior to Sleepwalking: Evidence of Interplay Between Sleep and Wakefulness. *Sleep*, 40(4):zsx024, April 2017. ISSN 1550-9109. doi: 10.1093/sleep/zsx024.

Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, 53(1):1–15, October 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.06.010.

Bradley J. Edelman, Nessa Johnson, Abbas Sohrabpour, Shanbao Tong, Nitish Thakor, and Bin He. Systems Neuroengineering: Understanding and Interacting with the Brain. *Engineering (Beijing)*, 1(3):292–308, September 2015. ISSN 2095-8099. doi: 10.15302/j-eng-2015078.

Edelman Robert R and Warach Steven. Magnetic Resonance Imaging. *New England Journal of Medicine*, 328(10):708–716, 1993. doi: 10.1056/NEJM199303113281008.

Jos J. Eggermont. The Correlative Brain. In Jos J. Eggermont, editor, *The Correlative Brain: Theory and Experiment in Neural Interaction*, pages 267–281. Springer, Berlin, Heidelberg, 1990. ISBN 978-3-642-51033-5. doi: 10.1007/978-3-642-51033-5_15.

Denis A Engemann, Oleh Kozynets, David Sabbagh, Guillaume Lemaître, Gael Varoquaux, Franziskus Liem, and Alexandre Gramfort. Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *eLife*, 9:e54055, May 2020. ISSN 2050-084X. doi: 10.7554/eLife.54055.

Denis A. Engemann, Apolline Mellot, Richard Höchenberger, Hubert Banville, David Sabbagh, Lukas Gemein, Tonio Ball, and Alexandre Gramfort. A reusable benchmark of brain-age prediction from M/EEG resting-state signals. *NeuroImage*, 262:119521, November 2022. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2022.119521.

Bruce Fischl. FreeSurfer. *Neuroimage*, 62(2):774–781, August 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2012.01.021.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously, December 2019.

R. A. Fisher. Design of Experiments. *Br Med J*, 1(3923):554, March 1936. ISSN 0007-1447.

R. A. Fisher. Statistical Methods for Research Workers. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 66–70. Springer, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_6.

Grant Fleming. How and Why to Interpret Black Box Models, March 2020.

Katja Franke and Christian Gaser. Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained? *Front Neurol*, 10: 789, August 2019. ISSN 1664-2295. doi: 10.3389/fneur.2019.00789.

Karl J. Friston. Modalities, modes, and models in functional neuroimaging. *Science*, 326(5951):399–403, October 2009. ISSN 1095-9203. doi: 10.1126/science.1174521.

Anna Fry, Thomas J. Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E. Allen. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol*, 186(9):1026–1034, November 2017. ISSN 1476-6256. doi: 10.1093/aje/kwx246.

Francis Galton. Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886. ISSN 0959-5295. doi: 10.2307/2841583.

Si Gao, Brian Donohue, Kathryn S. Hatch, Shuo Chen, Tianzhou Ma, Yizhou Ma, Mark D. Kvarta, Heather Bruce, Bhim M. Adhikari, Neda Jahanshad, Paul M. Thompson, John Blangero, L. Elliot Hong, Sarah E. Medland, Habib Ganjgahi, Thomas E. Nichols, and Peter Kochunov. Comparing empirical kinship derived heritability for imaging genetics traits in the UK biobank and human connectome project. *NeuroImage*, 245:118700, December 2021. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2021.118700.

Yue Gao, Abby Stevens, Rebecca Willet, and Garvesh Raskutti. Lazy Estimation of Variable Importance for Large Neural Networks, July 2022.

Pilar Garcés, David López-Sanz, Fernando Maestú, and Ernesto Pereda. Choice of Magnetometers and Gradiometers after Signal Space Separation. *Sensors (Basel)*, 17(12):2926, December 2017. ISSN 1424-8220. doi: 10.3390/s17122926.

Morgan Gautherot, Grégory Kuchcinski, Cécile Bordier, Adeline Rollin Sillaire, Xavier Delbeuck, Mélanie Leroy, Xavier Leclerc, Jean-Pierre Pruvo, Florence Pasquier, and Renaud Lopes. Longitudinal Analysis of Brain-Predicted Age in Amnestic and Non-amnestic Sporadic Early-Onset Alzheimer's Disease. *Front Aging Neurosci*, 13:729635, November 2021. ISSN 1663-4365. doi: 10.3389/fnagi.2021.729635.

Lukas A. W. Gemein, Robin T. Schirrmeister, Patryk Chrabąszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. Machine-learning-based diagnostics of EEG pathology. *NeuroImage*, 220: 117021, October 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.117021.

Joseph Giorgio, William J. Jagust, Suzanne Baker, Susan M. Landau, Peter Tino, Zoe Kourtzi, and Alzheimer's Disease Neuroimaging Initiative. A robust and interpretable machine learning approach using multimodal biological data to predict future pathological tau accumulation. *Nat Commun*, 13(1):1887, April 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28795-7.

Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90:15–35, October 2015. doi: 10.1016/j.csda.2015.04.002.

Matti Hämäläinen, Riitta Hari, Risto Ilmoniemi, Jukka Knuutila, and Olli Lounasmaa. Magnetoencephalography: Theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.*, 65:413, April 1993. doi: 10.1103/RevModPhys.65.413.

Brian Hepburn and Hanne Andersen. Scientific Method. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2021 edition, 2021.

Suzana Herculano-Houzel. The Human Brain in Numbers: A Linearly Scaled-up Primate Brain. *Front Hum Neurosci*, 3:31, November 2009. ISSN 1662-5161. doi: 10.3389/neuro.09.031.2009.

Oliver Hines, Karla Diaz-Ordaz, and Stijn Vansteelandt. Variable importance measures for heterogeneous causal effects, October 2023.

Simon M. Hofmann, Frauke Beyer, Sebastian Lapuschkin, Ole Goltermann, Markus Loeffler, Klaus-Robert Müller, Arno Villringer, Wojciech Samek, and A. Veronica Witte. Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain. *Neuroimage*, 261:119504, November 2022. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2022.119504.

Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted Permutation forces Extrapolation: Variable Importance Requires at least One More Model, or There Is No Free Variable Importance, October 2021.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A Benchmark for Interpretability Methods in Deep Neural Networks, November 2019.

Susan Hrisos, Martin P. Eccles, Jill J. Francis, Heather O. Dickinson, Eileen FS Kaner, Fiona Beyer, and Marie Johnston. Are there valid proxy measures of clinical behaviour? a systematic review. *Implementation Sci*, 4(1):37, July 2009. ISSN 1748-5908. doi: 10.1186/1748-5908-4-37.

Jui-Long Hung, Kerry Rice, Jennifer Kepka, and Juan Yang. Improving predictive power through deep learning analysis of K-12 online student behaviors and discussion board content. *Information Discovery and Delivery*, 48(4):199–212, January 2020. ISSN 2398-6247. doi: 10.1108/IDD-02-2020-0019.

R. Iniesta, D. Stahl, and P. McGuffin. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychol Med*, 46(12):2455–2465, September 2016. ISSN 1469-8978. doi: 10.1017/S0033291716001367.

Laurent Itti, Christof Koch, and Ernst Niebur. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20:1254–1259, December 1998. doi: 10.1109/34.730558.

J. Jaeger, R. Sengupta, and W. L. Ruzzo. Improved gene selection for classification of microarrays. *Pac Symp Biocomput*, pages 53–64, 2003. ISSN 2335-6928. doi: 10.1142/9789812776303_0006.

Mortaza Jamshidian, Robert I. Jennrich, and Wei Liu. A study of partial *F* tests for multiple linear regression models. *Computational Statistics & Data Analysis*, 51(12):6269–6284, August 2007. ISSN 0167-9473. doi: 10.1016/j.csda.2007.01.015.

Silke Janitza, Ender Celik, and Anne-Laure Boulesteix. A computationally fast variable importance test for random forests for high-dimensional data. *Adv Data Anal Classif*, 12(4):885–915, December 2018. ISSN 1862-5355. doi: 10.1007/s11634-016-0276-4.

Stephen C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, September 1967. ISSN 1860-0980. doi: 10.1007/BF02289588.

B. A. Jonsson, G. Bjornsdottir, T. E. Thorgeirsson, L. M. Ellingsen, G. Bragi Walters, D. F. Gudbjartsson, H. Stefansson, K. Stefansson, and M. O. Ulfarsson.

Brain age prediction using deep learning uncovers associated sequence variants. *Nat Commun*, 10(1):5409, November 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-13163-9.

Martin Jullum, Annabelle Redelmeier, and Kjersti Aas. groupShapley: Efficient prediction explanation with Shapley values for feature groups, June 2021.

Teresa M. Karrer, Danielle S. Bassett, Birgit Derntl, Oliver Gruber, André Aleman, Renaud Jardri, Angela R. Laird, Peter T. Fox, Simon B. Eickhoff, Olivier Grisel, Gaël Varoquaux, Bertrand Thirion, and Danilo Bzdok. Brain-based ranking of cognitive domains to predict schizophrenia. *Hum Brain Mapp*, 40 (15):4487–4507, October 2019. ISSN 1097-0193. doi: 10.1002/hbm.24716.

Leonard Kaufman and Peter Rousseeuw. *Finding Groups in Data: An Introduction To Cluster Analysis*. January 1990. ISBN 978-0-471-87876-6. doi: 10.2307/2532178.

Katherine A. Knutson and Wei Pan. Integrating brain imaging endophenotypes with GWAS for Alzheimer's disease. *Quant Biol*, August 2020. ISSN 2095-4697. doi: 10.1007/s40484-020-0202-9.

Peter Kochunov, Meghann C. Ryan, Qifan Yang, Kathryn S. Hatch, Alyssa Zhu, Sophia I. Thomopoulos, Neda Jahanshad, Lianne Schmaal, Paul M. Thompson, Shuo Chen, Xiaoming Du, Bhim M. Adhikari, Heather Bruce, Stephanie Hare, Eric L. Goldwaser, Mark D. Kvarta, Thomas E. Nichols, and L. Elliot Hong. Comparison of regional brain deficit patterns in common psychiatric and neurological disorders as revealed by big data. *NeuroImage: Clinical*, 29: 102574, January 2021. ISSN 2213-1582. doi: 10.1016/j.nicl.2021.102574.

Arinbjörn Kolbeinsson, Sarah Filippi, Yannis Panagakis, Paul M. Matthews, Paul Elliott, Abbas Dehghan, and Ioanna Tzoulaki. Accelerated MRI-predicted brain ageing and its associations with cardiometabolic and brain disorders. *Sci Rep*, 10(1):19940, November 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-76518-z.

Z. J. Koles, M. S. Lazar, and S. Z. Zhou. Spatial patterns underlying population differences in the background EEG. *Brain Topogr*, 2(4):275–284, 1990. ISSN 0896-0267. doi: 10.1007/BF01129656.

Padmavathi Kora, K. Meenakshi, K. Swaraja, A. Rajani, and Mantena Satyanarayana Raju. EEG based interpretation of human brain activity during yoga and meditation using machine learning: A systematic review. *Complementary Therapies in Clinical Practice*, 43:101329, May 2021. ISSN 1744-3881. doi: 10.1016/j.ctcp.2021.101329.

Nikolaos Koutsouleris, Christos Davatzikos, Stefan Borgwardt, Christian Gaser, Ronald Bottlender, Thomas Frodl, Peter Falkai, Anita Riecher-Rössler, Hans-Jürgen Möller, Maximilian Reiser, Christos Pantelis, and Eva Meisenzahl. Accelerated brain aging in schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders. *Schizophr Bull*, 40(5):1140–1153, September 2014. ISSN 1745-1701. doi: 10.1093/schbul/sbt142.

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures, June 2020.

P. C. Lauterbur. Image formation by induced local interactions. Examples employing nuclear magnetic resonance. 1973. *Clin Orthop Relat Res*, (244):3–6, July 1989. ISSN 0009-921X.

Kyubin Lee, Akshay Sood, and Mark Craven. Understanding Learned Models by Identifying Important Features at the Right Resolution, November 2018.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111, July 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1307116.

Herve Lemaitre, Aaron L. Goldman, Fabio Sambataro, Beth A. Verchinski, Andreas Meyer-Lindenberg, Daniel R. Weinberger, and Venkata S. Mattay. Normal age-related brain morphometric changes: Nonuniformity across cortical thickness, surface area and gray matter volume? *Neurobiol Aging*, 33(3): 617.e1–9, March 2012. ISSN 1558-1497. doi: 10.1016/j.neurobiolaging.2010.07.013.

Ling Li. The Differences among Eyes-Closed, Eyes-Open and Attention States: An EEG Study. In *2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM)*, pages 1–4, September 2010. doi: 10.1109/WICOM.2010.5600726.

Franziskus Liem, Gaël Varoquaux, Jana Kynast, Frauke Beyer, Shahrzad Kharabian Masouleh, Julia M. Huntenburg, Leonie Lampe, Mehdi Rahim, Alexandre Abraham, R. Cameron Craddock, Steffi Riedel-Heller, Tobias Luck, Markus Loeffler, Matthias L. Schroeter, Anja Veronica Witte, Arno Villringer, and Daniel S. Margulies. Predicting brain-age from multimodal imaging data captures cognitive impairment. *Neuroimage*, 148:179–188, March 2017. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2016.11.005.

Thomas J. Littlejohns, Jo Holliday, Lorna M. Gibson, Steve Garratt, Niels Oesingmann, Fidel Alfaro-Almagro, Jimmy D. Bell, Chris Boultwood, Rory Collins,

Megan C. Conroy, Nicola Crabtree, Nicola Doherty, Alejandro F. Frangi, Nicholas C. Harvey, Paul Leeson, Karla L. Miller, Stefan Neubauer, Steffen E. Petersen, Jonathan Sellors, Simon Sheard, Stephen M. Smith, Cathie L. M. Sudlow, Paul M. Matthews, and Naomi E. Allen. The UK Biobank imaging enhancement of 100,000 participants: Rationale, data collection, management and future directions. *Nat Commun*, 11(1):2624, May 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15948-9.

Chen-yu Liu, Arnab Maity, Xihong Lin, Robert O Wright, and David C Christiani. Design and analysis issues in gene and environment studies. *Environ Health*, 11:93, December 2012. ISSN 1476-069X. doi: 10.1186/1476-069X-11-93.

Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and Powerful Conditional Randomization Testing via Distillation. *arXiv:2006.03980 [stat]*, June 2021.

Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Yang Young Lu, Yingying Fan, Jinchi Lv, and William Stafford Noble. DeepPINK: Reproducible feature selection in deep neural networks. *arXiv:1809.01185 [cs, stat]*, September 2018.

Steven Luck. *An Introduction to The Event-Related Potential Technique*. January 2005.

Ian Lundberg, Jennie E. Brand, and Nanum Jeon. Researcher reasoning meets computational capacity: Machine learning for social science. *Social Science Research*, 108:102807, November 2022. ISSN 0049-089X. doi: 10.1016/j.ssresearch.2022.102807.

Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions, November 2017.

Hernán MA and Robins JM. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC., 2020.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5.1, pages 281–298. University of California Press, January 1967.

Christopher R. Madan and Elizabeth A. Kensinger. Predicting age from cortical structure across the lifespan. *Eur J Neurosci*, 47(5):399–416, March 2018. ISSN 0953-816X. doi: 10.1111/ejn.13835.

James D. Malley, Karen G. Malley, and Sinisa Pajevic. *Statistical Learning for Biomedical Data*. Cambridge University Press, February 2011. ISBN 978-1-139-49685-8.

T. McKelvey, Muhammad Ahmad, Ankur Teredesai, and Carly Eckert. Interpretable Machine Learning in Healthcare. August 2018.

Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux. Dictionary Learning for Massive Matrix Factorization.

Xinlei Mi, Baiming Zou, Fei Zou, and Jianhua Hu. Permutation-based identification of important biomarkers for complex diseases via machine learning models. *Nat Commun*, 12(1):3008, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22756-2.

Christine Michel, Manuela Stets, Eugenio Parise, Vincent M. Reid, Tricia Striano, and Stefanie Hoehl. Theta- and alpha-band EEG activity in response to eye gaze cues in early infancy. *NeuroImage*, 118:576–583, September 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.06.042.

Karla L. Miller, Fidel Alfaro-Almagro, Neal K. Bangerter, David L. Thomas, Essa Yacoub, Junqian Xu, Andreas J. Bartsch, Saad Jbabdi, Stamatios N. Sotiropoulos, Jesper L. R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W. Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M. Matthews, and Stephen M. Smith. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*, 19(11):1523–1536, November 2016. ISSN 1546-1726. doi: 10.1038/nn.4393.

Christoph Molnar. *Interpretable Machine Learning*. July 2022.

Christoph Molnar, Gunnar König, Bernd Bischl, and Giuseppe Casalicchio. Model-agnostic Feature Importance and Effects with Dependent Features – A Conditional Subgroup Approach, June 2021a.

Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models, August 2021b.

Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms, September 2011.

D. G. Murphy, C. DeCarli, M. B. Schapiro, S. I. Rapoport, and B. Horwitz. Age-related differences in volumes of subcortical nuclei, brain matter, and cerebrospinal fluid in healthy men as measured with magnetic resonance imaging. *Arch Neurol*, 49(8):839–845, August 1992. ISSN 0003-9942. doi: 10.1001/archneur.1992.00530320063013.

Julian Mutz and Cathryn M. Lewis. Lifetime depression and age-related changes in body composition, cardiovascular function, grip strength and lung function: Sex-specific analyses in the UK Biobank. *Aging (Albany NY)*, 13 (13):17038–17079, July 2021. ISSN 1945-4589. doi: 10.18632/aging.203275.

J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 0035-9238. doi: 10.2307/2344614.

Danielle Newby, Laura Winchester, William Sproviero, Marco Fernandes, Dai Wang, Andrey Kormilitzin, Lenore J. Launer, and Alejo J. Nevado-Holgado. Associations Between Brain Volumes and Cognitive Tests with Hypertensive Burden in UK Biobank. *J Alzheimers Dis*, 84(3):1373–1389, 2021. ISSN 1875-8908. doi: 10.3233/JAD-210512.

J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231: 289–337, 1933. ISSN 0264-3952.

Binh T. Nguyen, Bertrand Thirion, and Sylvain Arlot. A Conditional Randomization Test for Sparse Logistic Regression in High-Dimension, May 2022.

Tuan-Binh Nguyen, Jérôme-Alexis Chevalier, Bertrand Thirion, and Sylvain Arlot. Aggregation of Multiple Knockoffs. *arXiv:2002.09269 [math, stat]*, June 2020.

Kristin K. Nicodemus, James D. Malley, Carolin Strobl, and Andreas Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11(1):110, February 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-110.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG Data Corpus. *Frontiers in Neuroscience*, 10, 2016. ISSN 1662-453X.

S. Ogawa, T. M. Lee, A. S. Nayak, and P. Glynn. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn Reson Med*, 14(1):68–78, April 1990. ISSN 0740-3194. doi: 10.1002/mrm.1910140108.

R. Oostenveld and P. Praamstra. The five percent electrode system for high-resolution EEG and ERP measurements. *Clin Neurophysiol*, 112(4):713–719, April 2001. ISSN 1388-2457. doi: 10.1016/s1388-2457(00)00527-7.

Brian Patenaude, Stephen M. Smith, David N. Kennedy, and Mark Jenkinson. A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3):907–922, June 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.02.046.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12 (85):2825–2830, 2011. ISSN 1533-7928.

Matthieu Perrot, Denis Rivière, Alan Tucholka, and Jean-François Mangin. Joint Bayesian cortical sulci recognition and spatial normalization. *Inf Process Med Imaging*, 21:176–187, 2009. ISSN 1011-2499. doi: 10.1007/978-3-642-02498-6_15.

Sebastian G. Popescu, Ben Glocker, David J. Sharp, and James H. Cole. Local Brain-Age: A U-Net Model. *Front Aging Neurosci*, 13:761954, 2021. ISSN 1663-4365. doi: 10.3389/fnagi.2021.761954.

Mehdi Rahim, Bertrand Thirion, Alexandre Abraham, Michael Eickenberg, Elvis Dohmatob, Claude Comtat, and Gael Varoquaux. Integrating Multimodal Priors in Predictive Models for the Functional Characterization of Alzheimer's Disease. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 207–214, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24553-9. doi: 10.1007/978-3-319-24553-9_26.

Tim Räz. ML interpretability: Simple isn't easy. *Studies in History and Philosophy of Science*, 103:159–167, February 2024. ISSN 0039-3681. doi: 10.1016/j.shpsa.2023.12.007.

M Reddy, Vivekananda Makara, and Satish R U V N. Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering. 5, October 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, August 2016.

Andrei-Claudiu Roibu, Stanislaw Adaszewski, Torsten Schindler, Stephen M. Smith, Ana I.L. Namburete, and Frederik J. Lange. Brain Ages Derived from Different MRI Modalities are Associated with Distinct Biological Phenotypes. In *2023 10th IEEE Swiss Conference on Data Science (SDS)*, pages 17–25, June 2023. doi: 10.1109/SDS57534.2023.00010.

Jessica K. Roydhouse, Matthew L. Cohen, Henrik R. Eshoj, Nadia Corsini, Emre Yucel, Claudia Rutherford, Katarzyna Wac, Allan Berrocal, Alyssa Lanzi, Cindy Nowinski, Natasha Roberts, Angelos P. Kassianos, Veronique Sebille, Madeleine T. King, Rebecca Mercieca-Bebber, and ISOQOL Proxy Task Force and the ISOQOL Board of Directors. The use of proxies and proxy-reported measures: A report of the international society for quality of life research (ISOQOL) proxy task force. *Qual Life Res*, 31(2):317–327, February 2022. ISSN 1573-2649. doi: 10.1007/s11136-021-02937-8.

David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A. Engemann. Predictive regression modeling with MEG/EEG: From source power to signals and cognitive states. *NeuroImage*, 222:116893, November 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.116893.

David Sabbagh, Jérôme Cartailler, Cyril Touchard, Jona Joachim, Alexandre Mebazaa, Fabrice Vallée, Étienne Gayat, Alexandre Gramfort, and Denis A. Engemann. Repurposing electroencephalogram monitoring of general anaesthesia for building biomarkers of brain ageing: An exploratory study. *BJA Open*, 7, September 2023. ISSN 2772-6096. doi: 10.1016/j.bjao.2023.100145.

Debopam Samanta. Recent developments in stereo electroencephalography monitoring for epilepsy surgery. *Epilepsy Behav*, 135:108914, October 2022. ISSN 1525-5069. doi: 10.1016/j.yebeh.2022.108914.

Marc-Andre Schulz, B. T. Thomas Yeo, Joshua T. Vogelstein, Janaina Mourao-Miranada, Jakob N. Kather, Konrad Kording, Blake Richards, and Danilo Bzdok. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat Commun*, 11(1):4238, August 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18037-z.

Konstantinos Sechidis, Matthias Kormaksson, and David Ohlssen. Using knockoffs for controlled predictive biomarker identification. *Statistics in Medicine*, 40(25):5453–5473, 2021. ISSN 1097-0258. doi: 10.1002/sim.9134.

Meredith A. Shafto, Lorraine K. Tyler, Marie Dixon, Jason R. Taylor, James B. Rowe, Rhodri Cusack, Andrew J. Calder, William D. Marslen-Wilson, John Duncan, Tim Dalgleish, Richard N. Henson, Carol Brayne, Fiona E. Matthews,

and Cam-CAN. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: A cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurol*, 14:204, October 2014. ISSN 1471-2377. doi: 10.1186/s12883-014-0204-1.

Zack Y. Shan and Jim Lagopoulos. Precision Medicine for Brain Disorders: New and Emerging Approaches. *J Pers Med*, 13(5):872, May 2023. ISSN 2075-4426. doi: 10.3390/jpm13050872.

Lloyd S. Shapley. A Value for N-Person Games. Technical report, RAND Corporation, March 1952.

Janet Siebert. Integrated biomarker discovery: Combining heterogeneous data. *Bioanalysis*, 3(21):2369–2372, November 2011. ISSN 1757-6180. doi: 10.4155/bio.11.229.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014.

Nalini M. Singh, Jordan B. Harrod, Sandya Subramanian, Mitchell Robinson, Ken Chang, Suheyla Cetin-Karayumak, Adrian Vasile Dalca, Simon Eickhoff, Michael Fox, Loraine Franke, Polina Golland, Daniel Haehn, Juan Eugenio Iglesias, Lauren J. O'Donnell, Yangming Ou, Yogesh Rathi, Shan H. Siddiqi, Haoqi Sun, M. Brandon Westover, Susan Whitfield-Gabrieli, and Randy L. Gollub. How Machine Learning is Powering Neuroimaging to Improve Brain Health. *Neuroinform*, 20(4):943–964, October 2022. ISSN 1559-0089. doi: 10.1007/s12021-022-09572-9.

Stephen M. Smith. Fast robust automated brain extraction. *Hum Brain Mapp*, 17(3):143–155, November 2002. ISSN 1065-9471. doi: 10.1002/hbm.10062.

Stephen M. Smith and Thomas E. Nichols. Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron*, 97(2):263–268, January 2018. ISSN 1097-4199. doi: 10.1016/j.neuron.2017.12.018.

Stephen M. Smith, Yongyue Zhang, Mark Jenkinson, Jacqueline Chen, P. M. Matthews, Antonio Federico, and Nicola De Stefano. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage*, 17(1):479–489, September 2002. ISSN 1053-8119. doi: 10.1006/nimg.2002.1040.

Stephen M. Smith, Mark Jenkinson, Heidi Johansen-Berg, Daniel Rueckert, Thomas E. Nichols, Clare E. Mackay, Kate E. Watkins, Olga Ciccarelli, M. Zaheer Cader, Paul M. Matthews, and Timothy E. J. Behrens. Tract-based spa-

tial statistics: Voxelwise analysis of multi-subject diffusion data. *Neuroimage*, 31(4):1487–1505, July 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.02.024.

Stephen M. Smith, Karla L. Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph D. Ramsey, and Mark W. Woolrich. Network modelling methods for FMRI. *Neuroimage*, 54 (2):875–891, January 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.08.063.

Stephen M. Smith, Thomas E. Nichols, Diego Vidaurre, Anderson M. Winkler, Timothy E. J. Behrens, Matthew F. Glasser, Kamil Ugurbil, Deanna M. Barch, David C. Van Essen, and Karla L. Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat Neurosci*, 18(11):1565–1567, November 2015. ISSN 1546-1726. doi: 10.1038/nn.4125.

Stephen M. Smith, Diego Vidaurre, Fidel Alfaro-Almagro, Thomas E. Nichols, and Karla L. Miller. Estimation of brain age delta from brain imaging. *NeuroImage*, 200:528–539, October 2019. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.06.017.

Daichi Sone and Iman Beheshti. Neuroimaging-Based Brain Age Estimation: A Promising Personalized Biomarker in Neuropsychiatry. *Journal of Personalized Medicine*, 12(11), November 2022. doi: 10.3390/jpm12111850.

Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D. Gosling, Gabriella M. Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, Heinrich Hussmann, Bernd Bischl, and Markus Bühner. Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30): 17680–17687, July 2020. doi: 10.1073/pnas.1920484117.

Stephen M. Stigler. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, 1999. ISBN 978-0-674-83601-3. doi: 10.2307/j.ctv1pdrpsj.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, July 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-307.

Erik Štrumbelj and Igor Kononenko. An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research*, 11(1): 1–18, 2010. ISSN 1533-7928.

Michał Strzelecki and Pawel Badura. Machine Learning for Biomedical Application. *Applied Sciences*, 12(4):2022, January 2022. ISSN 2076-3417. doi: 10.3390/app12042022.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. 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):e1001779, March 2015. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001779.

Antonio Sutera, Gilles Louppe, Van Anh Huynh-Thu, Louis Wehenkel, and Pierre Geurts. From global to local MDI variable importances for random forests and when they are Shapley values. *arXiv:2111.02218 [cs, stat]*, November 2021.

Woo-Suk Tae, Byung-Joo Ham, Sung-Bom Pyun, Shin-Hyuk Kang, and Byung-Jo Kim. Current Clinical Applications of Diffusion-Tensor Imaging in Neurological Disorders. *J Clin Neurol*, 14(2):129–140, April 2018. ISSN 1738-6586. doi: 10.3988/jcn.2018.14.2.129.

S. Taulu, J. Simola, and M. Kajola. Applications of the signal space separation method. *IEEE Trans. Signal Process.*, 53(9):3359–3372, September 2005. ISSN 1053-587X. doi: 10.1109/TSP.2005.853302.

Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, Cam-CAN, and Richard N. Henson. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 144:262–269, January 2017. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.09.018.

Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, June 2015. doi: 10.1073/pnas.1507583112.

T. Teissier, E. Boulanger, and V. Deramecourt. Normal ageing of the brain: Histological and biological aspects. *Rev Neurol (Paris)*, 176(9):649–660, November 2020. ISSN 0035-3787. doi: 10.1016/j.neurol.2020.03.017.

Madhav Thambisetty, Jing Wan, Aaron Carass, Yang An, Jerry L. Prince, and Susan M. Resnick. Longitudinal changes in cortical thickness associated with normal aging. *Neuroimage*, 52(4):1215–1223, October 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.04.258.

N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289, January 2002. ISSN 1053-8119. doi: 10.1006/nimg.2001.0978.

Simon Valentin, Maximilian Harkotte, and Tzvetan Popov. Interpreting neural decoding models using grouped model reliance. *PLoS Computational Biology*, 16(1), January 2020. doi: 10.1371/journal.pcbi.1007148.

Gaël Varoquaux, Flore Baronnet, Andreas Kleinschmidt, Pierre Fillard, and Bertrand Thirion. Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. *Med Image Comput Comput Assist Interv*, 13(Pt 1):200–208, 2010. doi: 10.1007/978-3-642-15705-9_25.

David S. Watson and Marvin N. Wright. Testing conditional independence in supervised learning algorithms. *Mach Learn*, 110(8):2107–2129, August 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-06030-6.

Marie Wehenkel, Antonio Sutera, Christine Bastin, Pierre Geurts, and Christophe Phillips. Random Forests Based Group Importance Scores and Their Statistical Interpretation: Application for Alzheimer's Disease. *Frontiers in Neuroscience*, 12, 2018. ISSN 1662-453X.

Sebastian Weichwald and Jonas Peters. Causality in cognitive neuroscience: Concepts, challenges, and distributional robustness. https://arxiv.org/abs/2002.06060v2, February 2020.

Sebastian Weichwald, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. https://arxiv.org/abs/1511.04780v1, November 2015.

Brian D. Williamson, Peter B. Gilbert, Noah R. Simon, and Marco Carone. A General Framework for Inference on Algorithm-Agnostic Variable Importance. *Journal of the American Statistical Association*, 0(0):1–14, November 2021. ISSN 0162-1459. doi: 10.1080/01621459.2021.2003200.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, January 1992. ISSN 0893-6080. doi: 10.1016/S0893-6080(05)80023-1.

Choong-Wan Woo, Luke J. Chang, Martin A. Lindquist, and Tor D. Wager. Building better biomarkers: Brain models in translational neuroimaging. *Nat Neurosci*, 20(3):365–377, February 2017. ISSN 1546-1726. doi: 10.1038/nn.4478.

Gui XUE, Chuansheng CHEN, Zhong-Lin LU, and Qi DONG. Brain Imaging Techniques and Their Applications in Decision-Making Research. *Xin Li Xue Bao*, 42(1):120–137, February 2010. ISSN 0439-755X. doi: 10.3724/SP.J.1041.2010.00120.

Yuzhe Yang, Yuan Yuan, Guo Zhang, Hao Wang, Ying-Cong Chen, Yingcheng Liu, Christopher G. Tarolli, Daniel Crepeau, Jan Bukartyk, Mithri R. Junna, Aleksandar Videnovic, Terry D. Ellis, Melissa C. Lipford, Ray Dorsey, and Dina Katabi. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nat Med*, 28(10):2207–2215, October 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01932-x.

Jieping Ye, Kewei Chen, Teresa Wu, Jing Li, Zheng Zhao, Rinkal Patel, Min Bae, Ravi Janardan, Huan Liu, Gene Alexander, and Eric Reiman. Heterogeneous data fusion for alzheimer's disease study. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 1025–1033, New York, NY, USA, August 2008. Association for Computing Machinery. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1402012.

Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging*, 20(1):45–57, January 2001. ISSN 0278-0062. doi: 10.1109/42.906424.

Xuebin Zhao, Hong Chen, Yingjie Wang, Weifu Li, Tieliang Gong, Yulong Wang, and Feng Zheng. Error-based Knockoffs Inference for Controlled Feature Selection, March 2022.

Qingsong Zhou, Hai Liang, Zhimin Lin, and Kele Xu. Multimodal Feature Fusion for Video Advertisements Tagging Via Stacking Ensemble, August 2021.

Alexander Zien, Nicole Krämer, Sören Sonnenburg, and Gunnar Rätsch. The Feature Importance Ranking Measure. In Wray Buntine, Marko Grobelnik, Dunja Mladenić, and John Shawe-Taylor, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 694–709, Berlin, Heidelberg, 2009. Springer. ISBN 978-3-642-04174-7. doi: 10.1007/978-3-642-04174-7_45.