
MULTI-MODALITY MACHINE LEARNING PREDICTION OF MENTAL HEALTH-ARCHETYPES IN A DEEP PHENOTYPED YOUNG AND REPRESENTATIVE SAMPLE.

MASTER THESIS IN BIOINFORMATICS
ANDRÉS BARRENA CALDERÓN

October 2024

Supervisors: Per Qvist, Jakob Grove
Bioinformatics Research Center (BiRC)
Faculty of Natural Sciences
Aarhus University

*Multi-modality machine learning prediction of mental health-archetypes
in a deep phenotyped young and representative sample.*, Master Thesis
in Bioinformatics © October 2024

This thesis explores the application of multi-modal machine learning in predicting mental health archetypes in a young and well-represented cohort. By integrating polygenic scores, plasma blood markers and brain imaging data we aimed to identify archetypal profiles associated with psychiatric risk and resilience. Using Random Forest models and feature selection techniques, we achieved significant predictive power in differentiating individuals at risk for mental disorders, particularly within the A1 archetype (associated with high neuroticism and emotional dysregulation) and A5 archetype (associated with emotional stability and resilience). Our results reveal distinct genetic and metabolic signatures that delineate risk and protective archetypes. This research is a first step in a broader European initiative to map psychiatric phenotypes using deep phenotyping techniques.

CONTENTS

Preface	5
1 INTRODUCTION	7
1.1 Genetics in psychiatric disorders	8
1.1.1 What are PGS?	8
1.1.2 Polygenic Scores in Epidemiologic Studies	10
1.2 Blood markers in Psychiatric Disorders	11
1.3 Brain Imaging in Psychiatric Disorders	13
1.4 Multimodal Models in Predicting Psychiatric Disorders	15
1.5 Study background	15
1.5.1 The Need for a Continuum View of Mental Health	15
1.5.2 Archetypes in Mental Health Research	16
1.5.3 Why Archetypal Analysis?	17
1.5.4 The Role of Multimodal Data in Archetype Identification	18
2 MATERIALS AND METHOD	19
2.1 Study population and data	19
2.1.1 Polygenic scores	20
2.1.2 Magnetic resonance imaging (MRI)	20
2.1.3 Metabolomics data generation	21
2.1.4 Proteomics data generation	22
2.2 Analysis	23
2.2.1 Linear Analysis using Ordinary Least Squares (OLS)	23
2.2.2 Assumptions of OLS Regression	24
2.2.3 Non-linear analysis using Random Forest Regression	24
2.2.4 Feature Selection with Recursive Feature Elimination (RFE)	25
2.2.5 Random Forest Regression Model	25
2.2.6 Model Interpretation with SHAP	26
2.2.7 Multi-modal Model using Random Forest Regression	27
3 RESULTS	31
3.1 Metabolomics	31
3.1.1 Linear Regression Analysis	31
3.1.2 Non-Linear Regression Analysis	32
3.1.3 Classification Evaluation: Threshold-Based Analysis	35
3.2 PGS	36
3.2.1 Linear Regression Analysis	37
3.2.2 Non-Linear Regression Analysis	38

3.2.3	Classification Evaluation: Threshold-Based Analysis	41
3.3	Multi-modal Model	43
3.3.1	Late fusion	43
3.3.2	Early fusion	45
3.3.3	Classification Evaluation: Threshold-Based Analysis	47
3.4	Blood markers proteomics	50
3.4.1	Linear Regression Analysis	50
3.5	Structural brain imaging measures	52
3.5.1	Linear Regression Analysis	52
4	DISCUSSION	55
4.1	Key Findings	55
4.2	Model Performance	56
4.3	Interpretation in Context	56
4.4	Strengths and Limitations	56
4.5	Future Directions	57
5	CONCLUSION	59
6	CODE AVAILABILITY	61
	REFERENCES	63

PREFACE

The rise of mental health disorders across Europe, particularly among young populations, has prompted a renewed focus on understanding the underlying biological mechanisms that drive psychiatric risk. This master's thesis represents my contribution to this growing body of research. Building on archetypal models developed by previous studies, my research integrates genetic, blood markers, and brain imaging data to predict mental health archetypes, thereby providing new insights into psychiatric risk factors. This work was carried out at Aarhus University, Denmark, under the guidance of Associate Professor Per Qvist and Professor Jakob Grove, and is part of the EU COST Action CA18106 project. I hope that the findings of this thesis will contribute to the ongoing efforts to improve mental health screening and interventions.