Predictive models and representations for neuroimaging and genetic data

A Dissertation Presented

by

Syed Fahad Sultan

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Computer Science

Stony Brook University

May 2022

Stony Brook University

The Graduate School

Syed Fahad Sultan

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Dr. Steven Skiena – Dissertation Advisor
Professor, Department of Computer Science

Dr. Andrew Schwartz – Chairperson of Defense
Associate Professor, Department of Computer Science

Dr. Stanley Bak
Assistant Professor, Department of Computer Science

Dr. Lilianne Mujica-Parodi
Professor, Biomedical Engineering Department
Stony Brook University

This dissertation is accepted by the Graduate School.

Eric Wertheimer
Dean of the Graduate School

Abstract of the Dissertation

Predictive models and representations for neuroimaging and genetic data

by

Syed Fahad Sultan

Doctor of Philosophy

in

Computer Science

Stony Brook University

2022

In this work, I present multiple computational models for extracting relevant features and learning predictive representations for neuroimaging and genetic data in UK-Biobank. These models are designed with the goal of providing improved diagnoses and a better understanding of medical conditions and other phenotypes. This includes a machine learning based predictive framework called Neuropredictome that identifies statistically significant linkages between 4928 phenotypes and neuroimaging features of 19,831 subjects. I also provide a novel quantitative method that uses deep learning based text embeddings to evaluate how well Neuropredictome's results align with 14,371 previously published peer-reviewed research articles. Next, I present a generalized framework based on state space systems that bridges the gap between network theory and control theory and extracts fMRI derived control circuits. This framework has the scalability required to mine mega-scale datasets, hirtherto not possible using existing methods. In a purely data-driven manner, without priors, I demonstrate that the framework

identifies thalamus-linked prefrontal-limbic and ventral stream subcircuits, selectively engaged during sensorimotor processing of affective and non-affective stimuli. I demonstrate that circuit-wide dysregulation, defined by degree of drift from healthy trajectories, tracks symptom severity for neuroticism, depression, and bipolar disorder. I also present methods for constructing low-dimensional vector representations (embeddings) of large-scale genotyping data, capable of reducing genotypes of hundreds of thousands of SNPs to 100-dimensional embeddings that retain substantial predictive power for inferring medical phenotypes. I also demonstrate how these genotype embeddings can be used for sharing sensitive medical data while preserving subject anonymity. Finally, using structural and functional neuroimaging in conjunction with cognitive tests, I show that type 2 diabetes mellitus accelerates brain aging and cognitive decline. Together, I believe such computational techniques can significantly advance modern medicine and treatment while enabling several scientific discoveries that revolutionize human health.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

How can three pounds of flesh still outperform the most powerful computers to date, and that too while consuming less energy than an average lightbulb? The inner working of the human brain undoubtedly remains one of the biggest scientific mysteries to date. The set of technologies that hold the greatest promise to ever solving this great mystery are known as neuroimaging. These technologies allow us to peek inside the human brain from the outside, and capture images that reveal not only what it is made up of but also what it is doing at any point in time.

One neuroimaging technology in particular stands out as having revolutionized our ability to image the human brain: Magnetic Resonance Imaging (MRI). MRI provides us the ability to safely watch the human brain in action, which in turn allows us to understand how the brain performs many of its functions. MRI scans can be grouped into two broad categories: structural and functional MRI, each measuring a distinct but equally important aspect of the brain. Structural MRI captures the makeup of brain's anatomy, such as how much water or fat is present in the tissue. Because different parts of the brain contain different amounts of water and fat, they show up on the MRI image as either brighter or darker. Structural MRIs of the brain are often used for detecting diseases and for understanding differences in size and shape of different brain regions between people, but they fail to capture what the brain is doing. For that, we need functional MRI (fMRI). fMRI picks up on the after effects of brain activity by sensing the changes in amount of oxygen in the blood, for a particular area, over a window of time.

In many cases fMRI can even allow us to decode what people are experiencing or thinking about by looking at their brain activity, when they are

perfoming a task or simply resting. MRI has also shown us how experiences change the brain, and how individual human brains change over time from childhood to old age. It has also shown us that all human brains follow the same basic principle, while also exhibiting key differences across people, thus giving insights into brain dysfunctions that relate to various mental illnesses.

fMRI relies upon a set of critical chemical and biological dominoes that all had to be perfectly in line for the technology to work. The first biological fact that makes fMRI possible is that the firing of neurons is relatively localized in the brain. For example, all humans, in fact most animals, possess a specific part of the brain, often in the back of the head, that processes visual information, generally called the visual cortex. This part of the brain responds to information of different parts of the visual world coming from the eyes. Similarly, different parts of the brain are responsible for doing different things. In other works, there is some degree of modularity in our brains. This modularity and localization of function ultimately allows us to *decode* from the brain scans, what the person in the scanner might have been doing or thinking of at the time of the scan.

Another aspect of the brain that is critical to fMRI being extremely effective as a tool for studying human brains is that brains are organized in a relatively consistent manner across individuals. All humans, for instance, in addition to the visual cortex, have a region of the brain, at the rear of the frontal lobe, that controls the movement of hands and feets. The alignment of different brain regions across people is not always perfect, in fact, far from it, but it's good enough to allow group level analyses.

The third crucial factor is that the localized firing of neurons result in increase of blood flow at that location. Without such localization, we wouldn't be able to narrow down the precise region of interest relevant to the neurons that caused the activity.

The final important characteristic is that blood flow response to an active region of the brain active neurons brings with it glucose and oxygen for the neurons. The amount of glucose supplied is just the right amount to account for the energy needed by the neurons to fire, but the amount of oxygen sent is in surplus, relative to the small amount actually needed by the neurons. It is precisely this overflow of oxygenated blood that lets the fMRI scanner detect the activity of neurons.

fMRI has ultimately raised some extremely fundamental questions about how we view ourselves as humans. If all cognition is just neuronal activations, that we can visualize with MRI, does that mean telepathy is only a short step away? Would we be able to read each other's minds? Have we solved the mystery of human consciousness? In what sense are we truly responsible

for our choices? Is addiction the same as any other physiological disease or a failure of self control or both?

It is these kinds of questions have resulted in recent years witnessing an explosion of interest in neuroimaging research. Each year, thousands of brain-imaging studies explore the links between brain and behavior. Last year, over 35,000 fMRI articles were published [2], including more than 6,000 publications containing the term "neuroimaging" [3]. These include study of a truly wide array of behaviors and characteristics, ranging from mental illnesses, cognitive acuity and adolescent development to left-handedness, gambling and political leanings.

However most neuroimaging research involves probing single variables to make narrow predefined discoveries, typically on a small data set. Individual studies therefore, seldom have statistical power to establish fully trustworthy results [4, 5]. Most neuroimaging studies suffer from the problem of small sample sizes. Median sample size of fMRI studies in 2015 was 28.5 subjects and 75th percentile of sample size in cognitive neuroscience journals published between 2011 to 2014 was 28 subjects. Working with such small sample sizes raises concerns regarding representativeness of the sample and inflated false discovery rate, particularly in its usage for clinical purposes, as well as regarding reproducibility. Low statistical power reduces the probability of detecting statistically significant results. As a consequence, in neuroimaging, effect sizes are often overestimated and reproducibility of results is fairly low. Finding consistent aggregate trends in the knowledge acquired across these studies is crucial but daunting.

Technological advances in recent decades however, now allow generating, collecting and analyzing massive data sets. This new scale in availability of data necessitates moving away from traditional methods that are used to infer statistically relevant effects in carefully chosen variables towards more machine learning based pattern-recognition algorithms that can identify relevant biological signatures in an unbiased way and produce results that generalize better to new unseen data.

1.2 Predictive models and representations

In this document I use an unprecedentedly large-scale neuroimaging and genetic data set UKBiobank [6] (N=19,831) and detail multiple computational models and frameworks for effectively predicting disease and other phenotypes, while also shedding critical light on the underlying mechanisms at play.

In Chapter 2, I present a novel prediction based framework called *Neuropredictome* that allows identification of statistically significant linkages be-

tween phenotypes and neuroimaging features from four different modalities. I evaluate phenotype linkage to brain activity of 19,831 subjects with 4926 variables, pertaining to the health, physiology, psychology, social and economic state. I corroborate my identified regions of the brain with previous work by providing a novel quantitative evaluation of how well my results align with existing meta-analyses of 14,371 published neuroimaging research articles. My results show that neuroimaging reveals as many neurological links to physical and lifestyle factors as to cognitive factors, supporting a more integrative approach to medicine that considers disease interactions between multiple organs and systems. My analysis is presented as a public resource at `https://neuropredictome.com` providing an interpretable view of human brain organization and decoding, to assist in hypothesis generation and evaluating future studies.

Mental health disorders are increasingly believed to be biological disorders involving some form of breakdown in brain circuitry. However, most existing work in computational neuroscience focuses on networks rather than circuits. Techniques that *do* allow modeling circuits in the brain at the macro scale unfortunately do not scale well enough to allow studying circuits at the whole-brain level. This not only forgoes recent advances in high-resolution non-invasive human neuroimaging technologies and availability of large data sets, but also severely limits the clinical applicability of such work. In Chapter 3, I bridge the gap between network theory and control theory and present a scalable framework for extracting fMRI-derived (generative) control circuits, then use circuit trajectories to estimate their control error. Using synthetic circuits, I first demonstrate that my framework accurately identifies each circuit's architecture and models its dynamics by estimation of transfer functions at the individual node level. As a use case, I then apply the framework to human task-based functional MRI data from UK Biobank (N=19,831). In a purely data-driven manner, without priors, my framework identified thalamus-linked *prefrontal-limbic* and *ventral stream* subcircuits, selectively engaged during sensorimotor processing of affective and non-affective stimuli. Finally, I demonstrate that circuit-wide dysregulation, defined by degree of drift from healthy trajectories, tracks symptom severity for neuroticism (*ventral subcircuit*), depression (*prefrontal-limbic subcircuit*), and bipolar disorder (*full circuit*).

In Chapter 4, I develop methods for constructing low-dimensional vector representations (embeddings) of large-scale genotyping data, capable of reducing genotypes of hundreds of thousands of SNPs to 100-dimensional embeddings that retain substantial predictive power for inferring medical phenotypes. I demonstrate that embedding-based models yield an average F-score of 0.605

on a test of ten phenoypes (including BMI prediction, genetic relatedness, and depression) versus 0.339 for baseline models. Genotype embeddings also hold promise for creating sharing data while preserving subject anonymity: I show that they retain substantial predictive power even after anonymization by adding Gaussian noise to each dimension.

Type 2 diabetes mellitus (T2DM) is known to be associated with neurobiological and cognitive deficits; however, their extent, overlap with aging effects, and the effectiveness of existing treatments in the context of the brain are currently unknown. In Chapter 5, I characterize neurocognitive effects independently associated with T2DM and age in a large cohort of human subjects from the UK Biobank with cross-sectional neuroimaging and cognitive data. I then proceeded to evaluate the extent of overlap between the effects related to T2DM and age by applying correlation measures to the independently characterized neurocognitive changes. My findings were complemented by meta-analyses of published reports with cognitive or neuroimaging measures for T2DM and healthy controls (HC). I also evaluated in a cohort of T2DM diagnosed individuals using UK Biobank how disease chronicity and metformin treatment may influence the characterized deficiencies. All analyses were Bonferroni corrected. Duration of T2DM ranged from 0–45 years (mean 9.7±7.9 years); 559 were treated with metformin alone, while 473 were unmedicated. My meta-analysis evaluated 34 cognitive studies (N=22,231) and 60 neuroimaging studies: 30 of T2DM (N=866) and 30 of aging (N=1088). As compared to age, sex, and education-matched HC, T2DM was associated with marked cognitive deficits, particularly in executive functioning and processing speed. Likewise, I found that the diagnosis of T2DM was significantly associated with gray matter atrophy and reorganization of brain activity. The structural and functional changes associated with T2DM show marked overlap with the effects correlating with age but appear earlier, with disease duration linked to more severe neurodegeneration. The neurocognitive impact of T2DM suggests marked acceleration of normal brain aging, by approximately 24%, made worse with chronicity. As such, neuroimaging-based biomarkers may provide a valuable adjunctive measure of T2DM progression and treatment efficacy based on neurological outcomes. Conventional T2DM treatments must be reevaluated with respect to restraining neurodegeneration.

Variables declared important by traditional null hypothesis significance testing can be incongruent with the variables that maximize predictive performance in new individuals or settings. This combined with small sample sizes is a major factor in the low reproducibility crisis [7] plaguing the scientific world today. The advent of "big data" in neuroscience and biomedicine in conjunction with with recent advances in machine learning and deep learning

presents an tremendous opportunity to address these problems. The data-rich neuroscientist and geneticist can ask many new questions that could probably never be addressed quantitatively before.

In this document, I present models and analytical techniques that leverage the availability of large-scale data sets made possible by new modes of data dissemination and open science and also use state of the art machine learning techniques better suited for their analyses. These not only optimize for predictions and reproducibility of results for new previously unseen subjects but also evaluate them against decades of past and existing research. Most importantly, the results presented here argue that prediction does not have to come at the cost of interpretability. Mechanistic insight does not need be compromised by optimizing predictive accuracy and the two goals can be achieved together and synergistically.

The remaining of the document is structured as follows:

1. **Chapter 2** - Building a machine learning based predictive framework called *Neuropredictome* that links 4928 phenotypes to brain features from fMRI scans of 19,831 subjects in UK-Biobank. The framework is then evaluated against 14,371 published peer-reviewed papers using deep learning based text embeddings.

2. **Chapter 3** - Modeling the brain as an amalgam of control circuits and modeling disease as the control circuit's failure to maintain homeostatic or allostatic control. Using a system of differential equations to model the complex causal relationships and feedback loops that constitute brain circuits, I show how the prefrontal-limbic circuit [8] and the cortico-thalamic circuit [9] can be discovered from task-fMRI scans in UK Biobank. Modeling prediction error as the measure of dysregulation, I also show how the two circuits are dysregulated across the following psychiatric disorders: neuroticism, depression and bipolar disorder.

3. **Chapter 4** - I develop methods for constructing low-dimensional vector representations (embeddings) of large-scale genotyping data, capable of reducing genotypes of hundreds of thousands of SNPs to 100-dimensional embeddings that retain substantial predictive power for inferring medical phenotypes. I also show that these embeddings retain substantial predictive power even after anonymization by adding Gaussian noise to each dimension.

4. **Chapter 5** - I characterize neurocognitive effects independently associated with Type 2 Diabetes (T2DM) and Aging in UK Biobank with cross-sectional neuroimaging and cognitive data. My results show a significant overlap in the effects related to T2DM and Aging, complemented by meta-analyses of published reports with cognitive or neuroimaging measures. I also evaluate how disease chronicity and metformin treatment may influence the characterized deficiencies. My results show T2DM accelerates normal brain aging, by approximately 24%, made worse with chronicity.