Honor's Thesis Proposal

Improving Parkinson's Disease Prediction By Leaveraging Disparate Datasets

XXXXX XXXXXX

July 5, 2019

1 Project Purpose

The purpose of this thesis is to improve Parkinson's Disease prediction using modern machine learning and statistical methods by leveraging many Parkinson's datasets that already exist.[8, 9, 10, 11]

2 Project Importance

Parkinson's Disease is something that is quite personal to me. My grandfather was diagnosed with a case of early-onset Parkinson's Disease. As his condition progressed, he experienced a wide variety of symptoms: constant tremors, occaisional hallucinations, and a dramatic loss of autonomy. As a result of Parkinson's Disease, my grandfather had to retire early from his professorship at California State University, be confined to a wheelchair for the last fifteen years of his life, and had extreme difficulty communicating orally.

Instead of attempting to determine a cause or a potential cause, the goal of this thesis is to use machine learning methods to predict a Parkinson's diagnosis, thereby helping patients to recieve treatment sooner and enjoy an improved quality of life.

3 Project Overview

Parkinson's Disease is a progressive nervous system condition that affects an estimated seven to ten million people worldwide. Unfortunately, research has not yet revealed a distinct cause or cure for this condition.[1, 2, 3, 4] Instead, most treatments available to Parkinson's patients alleviate symptoms and slow down the progression of the disease – including medications designed to increase neurotransmitter levels (such as MAO B inhibitors like selegiline, rasagiline and safinamide) and surgical procedures like deep brain stimulation.[7]

In general, the sooner Parkinson's disease is diagnosed, the more treatment options are available. Similarly, all other things being equal, a patient who is diagnosed earlier can expect to have a higher quality of life than patients diagnosed later. Thus, a broad body of Parkinson's research exists to improve diagnosis by predicting Parkinson's disease diagnoses.[12, 6]

A variety of datasets have been published that include genetic data, age, telemetry from accelerometers and other sensors, and more. Unfortunately, these datasets typically are considered to be "wide," that is, they have many features (sometimes in the hundreds) but comparitively few instances (sometimes also in the hundreds). From a machine learning and data science perspective, this can be problematic. Most modern

machine learning algorithms (such as neural networks) require high volume datasets in order to generalize well and prevent overfitting. Researchers have attempted to do their best in spite of these limitations.[5]

I hypothesize, however, that these limitations can be largely sidestepped by combining disparate datasets together. Each dataset shares features with other datasets, and could be augmented by identifying a valid regression model that predicts the values of features that are unique to a particular dataset by considering shared features. Datasets could further be augmented by leveraging a generative adversarial neural network. From there, the augmented data can be aggregated into a larger dataset. From there, dimensionality reduction techniques could be used to operate on a lower-dimensional manifold of the input space and be fed into a machine learning algorithm to produce results that are more accurate than those obtained by merely considering a single dataset.

4 Thesis Committee

My thesis committee will be comprised of the following professors:

- *Faculty Advisor:* Dr. Mark Clement has agreed to serve as the faculty advisor for this thesis. His research focuses on bioinformatics and computational biology, especially DNA sequence assembly algorithms. Dr. Clement has recently published research related to Alzheimer's Disease, which is closely related to Parkinson's Disease, and continues to push the boundaries of his field with his innovative lab. In addition to his extensive research experience, he provides mentorship for Computer Science students by serving as faculty undergraduate advisor.
- *Faculty Reader:* Dr. Tony Martinez has agreed to serve as the faculty reader for this thesis. Dr. Martinez's research has published over 150 papers on machine learning topics, including neural networks. Dr. Martinez's laboratory has recently published papers that aim to tackle similar problems as this thesis – such as learning on training data that is incomplete and leveraging methods to improve training accuracy of machine learning models.
- *Honors Coordinator:* Dr. Seth Holladay serves as the Computer Science Honors Coordinator and works closely with the Computer Science Department's award-winning animation program.

5 Project Timeline

- *Submit Thesis Proposal:* July 5th
- *Research:* July 5th to December 15th
- *Thesis Publication:* January 2nd to March 10th
- *Thesis Defense:* March 3rd
- *Thesis Poster:* March 17th

6 IRB/IACUC Approvals

This research is not going to involve human or animal subjects; approval from these university committees will not be required.

7 Funding

This research should not require any funding. There is no fee for accessing the databases I've identified for research, nor will I need funds for equipment or travel. I anticipate requiring resources from BYU's Office of Research Computing, but these resources are provided free of charge to the campus community.

8 Culminating Experience

As a child, I got very close with my grandpa. One day – with all the innocence that being a four year-old brings – I promised him that I would cure Parkinson's. Unfortunately, the research does not indicate that we are close to finding a cure. However, I consider discharging that promise to the best of my ability – helping to more quickly identify and slow the progression of the disease – to be the primary reward of this research.

At this time, I do not have any plans to submit my paper for publication or presentation at an outside conference, though I do plan on submitting my thesis for presentation at the 2020 Honor's Conference. Rather than present at a conference, my intent is to use this experience to prepare myself to study computer science as a graduate student.

9 References

[1] Mihael H. Polymeropoulos et al. *Mutation in the Alpha-Synuclein Gene Identified in Families with Parkinson's Disease.* June 1997. URL: https://science.sciencemag.org/content/276/5321/2045.

[2] A. B. Singleton et al. *Alpha-Synuclein Locus Triplication Causes Parkinson's Disease.* Oct. 2003. URL: https://science.sciencemag.org/content/302/5646/841.

[3] *Cloning of the Gene Containing Mutations that Cause PARK8-Linked Parkinson's Disease.* Nov. 2004. URL: https://www.sciencedirect.com/science/article/pii/S0896627304006890.

[4] Jose A Obeso et al. *Missing pieces in the Parkinson's disease puzzle.* June 2010. URL: https://www.ncbi.nlm.nih.gov/pubmed/20495568.

[5] *Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach.* Mar. 2013. URL: https://www.sciencedirect.com/science/article/abs/pii/S1746809413000359.

[6] Tarigoppula V. S. Sriram et al. *Diagnosis of Parkinson Disease Using Machine Learning and Data Mining Systems from Voice Dataset.* 2015. URL: https://link.springer.com/chapter/10.1007/978-3-319-11933-5_17.

[7] *Parkinson's disease.* June 2018. URL: https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/diagnosis-treatment/drc-20376062.

[8] URL: http://fcon_1000.projects.nitrc.org/indi/retro/parkinsons.html.

[9] URL: https://archive.ics.uci.edu/ml/datasets/parkinsons.

[10] URL: https://archive.ics.uci.edu/ml/datasets/Parkinson's%20Disease%20Classification.

[11] *Data Sets.* URL: https://www.michaeljfox.org/data-sets.

[12] *Predicting the Risk of Parkinson's Disease with a Mathematical Formula.* URL: https://www.michaeljfox.org/grant/predicting-risk-parkinsons-disease-mathematical-formula.