

**The Edward S. Rogers Sr. Department of
Electrical and Computer Engineering
University of Toronto**

ECE496Y Design Project Course - Group Final Report
Machine Learning on Brain Graph

Project number: 2018937

Supervisor: Dr. Ashish Khisti

Administrator / Section number: Mr. Ross Gillett / 07

Team members:

Shi Hu

Jixiong Deng

Yuxi Cai

Emails:

rock.hu@mail.utoronto.ca

jixiong.deng@mail.utoronto.ca

yuxijune.cai@mail.utoronto.ca

Date of Submission: March 21, 2019

Group Final Report Attribution Table

This table should be filled out to accurately reflect who contributed to each section of the report and what they contributed. Provide a **column** for each student, a **row** for each major section of the report, and the appropriate codes (e.g. 'RD, MR') in each of the necessary **cells** in the table. You may expand the table, inserting rows as needed, but you should not require more than two pages. The original completed, and signed form must be included in the hardcopies of the final report. Please make a copy of it for your own reference.

Section	Student Names		
	1. Jixiong Deng	2. Shi Hu	3. Yuxi Cai
Cover page, attribution table, Group Highlights and Individual Contributions	RS RD MR	RS RD MR	RS RD MR
Acknowledgements	RS RD MR	ET	MR
Executive Summary	RS	RS RD MR	RS RD MR
Introduction	RS RD MR	RS	MR
Final Design	RS	RS RD MR	RS RD MR
Testing and Verification	RS RD MR	RS	RS MR
Summary and Conclusions	RS	RS RD MR	RS MR
References	RD MR	RD MR	RD MR
Appendices	RD MR	RD MR	RD MR
All	FP CM ET	FP CM ET	FP CM ET

Abbreviation Codes:

Fill in abbreviations for roles for each of the required content elements. You do not have to fill in every cell. The "All" row refers to the complete report and should indicate who was responsible for the final compilation and final read through of the completed document.

RS – responsible for research of information

RD – wrote the first draft

MR – responsible for major revision

ET – edited for grammar, spelling, and expression

OR – other

"All" row abbreviations:

FP – final read through of complete document for flow and consistency

CM – responsible for compiling the elements into the complete document

OR - other




If you put OR (other) in a cell please put it in as OR1, OR2, etc. Explain briefly below the role referred to:

OR1: enter brief description here

OR2: enter brief description here

Signatures

By signing below, you verify that you have read the attribution table and agree that it accurately reflects your contribution to this document.

Name	Jixiong Deng	Signature		Date:	Mar 21, 2019
Name	Shi Hu	Signature		Date:	Mar 21, 2019
Name	Yuxi Cai	Signature		Date:	Mar 21, 2019
Name		Signature		Date:	

Voluntary Document Release Consent Form¹

To all ECE496 students:

To better help future students, we would like to provide examples that are drawn from excerpts of past student reports. The examples will be used to illustrate general communication principles as well as how the document guidelines can be applied to a variety of categories of design projects (e.g. electronics, computer, software, networking, research).

Any material chosen for the examples will be altered so that all names are removed. In addition, where possible, much of the technical details will also be removed so that the structure or presentation style are highlighted rather than the original technical content. These examples will be made available to students on the course website, and in general may be accessible by the public. The original reports will not be released but will be accessible only to the course instructors and administrative staff.

Participation is completely voluntary and students may refuse to participate or may withdraw their permission at any time. Reports will only be used with the signed consent of all team members. Participating will have no influence on the grading of your work and there is no penalty for not taking part.

If your group agrees to take part, please have all members sign the bottom of this form. The original completed and signed form should be included in the hardcopies of the final report.




Sincerely,
Khoman Phang
Phil Anderson
ECE496Y Course Coordinators

Consent Statement

We verify that we have read the above letter and are giving permission for the ECE496 course coordinator to use our reports as outlined above.

Team #: 2018937 Project Title: Machine Learning on Brain Graph

Supervisor: Ashish Khisti Administrator: Ross Gillett

Name <u>Jixiong Deng</u>	Signature <u></u>	Date: <u>Mar 21, 2019</u>
Name <u>Shi Hu</u>	Signature <u></u>	Date: <u>Mar 21, 2019</u>
Name <u>Yuxi Cai</u>	Signature <u></u>	Date: <u>Mar 21, 2019</u>
Name _____	Signature _____	Date: _____

¹ This form will be detached from the hardcopy of the final report. Please make sure you have nothing printed on the back page.

Executive Summary

In summary, this project is to provide a machine learning solution which takes functional magnetic resonance imaging (fMRI) as inputs and outputs proper labels to classify inputs in different types of brain diseases.

The human brain is one of the most complex networks known to man. The field of network neuroscience still remains in its very early stage while machine learning is rising and showing its power in recent years. Given machine learning on graph-structured data is relatively new and challenging, the team is motivated to apply machine learning techniques to brain graphs to explore the possibility of improvement in this area of study.

In general, our final design can be divided into two parts. First is the data science module, which converts raw fMRI signals from ADHD-200 and ABIDE databases to brain graphs. The dataset contains three types of label: healthy, attention deficit hyperactivity disorder (ADHD) and autism spectrum disorder (ASD). The module stores the resulted adjacency matrices and labels into matrix.csv and label.csv files. The second module consists of a machine learning model based on multi-layer perceptron (MLP) algorithm. The machine learning module is capable of assigning labels containing information of the type the brain disorder to subjects according to the brain graphs with validation accuracy of 93%. The model was decided based on the implementation and training of several different algorithm, including but not limited to Support Vector Machine (SVM) and Convolutional Neural Network (CNN). MLP was selected as the final solution because it provides the highest validation accuracy given the data output by the data science module.

In the final proposal, the team came up with a number of requirements, constraints and objective to ensure the quality of the final design. The final design satisfies all the requirement and is able to give predictions of diseases for the given brain graphs with an accuracy notably higher than the accuracy of random prediction. Further improvement on the dataset is needed to include preprocessing and the result will be presented in the design fair.

Group Highlights

In general, the progress has been delayed comparing to the original plan. Most of the progress was made in the middle of the fall semester and the beginning of the winter semester. The source of delay is two-folded: one being that we failed to take into account in the time we could spend on the project especially during the exam period in the fall semester; the other being that there was a change in the source of datasets.

The major change of our project is caused by the unavailability of preferred fMRI datasets. Since extra efforts were required to investigate more datasets, the data collection task has been delayed. However, we prioritized and created new tasks in order to proceed. For the data science module, instead of collecting data before investigating fMRI conversion tools, we tested tools like Nilearn with a smaller set of data. And the prototyping of conversion pipeline has been started before the collection of datasets is finalized. For the two machine learning modules, instead of adjusting the implementation based on actual adjacency matrices, the two modules were tested with images and synthetic brain graph datasets.

All team members have made a sufficient amount of effort into the project so that by the time of the final report, the system is in shape. However, more changes are still needed before the design fair. For the data science module, the scripts for `fetchData` and `toBrainGraph` are both completed and tested. However, because of insufficient communication with the supervisor, we made a wrong trade-off between preprocessing the data and the number of diseases and sample. As a result, at the time this report is written, the data used include more than 1000 samples of healthy control and two different diseases. However, as our supervisor advised, preprocessing data is more important than including multiple diseases in the data set. Therefore we will need to preprocess at least one group of the data before the design fair. For machine learning modules, a number of possible algorithms have been examined and the accuracy results are retrieved on time. However, as the datasets change, we will need to re-access those models to see whether the changes will affect the selection of models.

Individual Contributions

Yuxi Cai

The purpose of our project is to develop a machine learning solution to classify brain diseases using functional magnetic resonance imaging (fMRI) signals retrieved from public databases. The project is divided into two parts, one is the data science module, which transforms fMRIs into adjacency matrices, the other is the machine learning module, which is trained and evaluated with the output from the previous module. My main responsibility for this project is to find sources for, design and implement the data science module for our project.

To start with, I first configured my development environment and make sure the version of the tools are compatible. Since I have not learned about parallel programming before, I did some research on and learned about the library about parallel programming in Python.

After that, I started the data collecting process by investigating the public databases recommended by our supervisor. However, the release of several datasets from the Human Connectome Project, which was considered the best fit for our project in our proposal, has been delayed. Therefore, I looked into a list of other datasets and need confirmation from our supervisor. These changes have delayed this task significantly because retrieving data from different databases means there is a higher chance of inconsistency in the data collection process so more work was done to access the quality of the databases. Meanwhile, I decided to use resting-state fMRI, which presents information about brain activity when the subject is at rest. It is because resting-state fMRIs are widely available in those databases and there is less variance in the status of the subjects among different datasets. After researching the tools for converting fMRI to adjacency matrices, I decided to use the Nilearn package because the interface is easy to understand and it has been used in a number of published studies.

By the time of the report, the data used includes more than 1000 samples of healthy control and two different diseases. However, as our supervisor advised, preprocessing data is more important than including multiple diseases in the data set. I will need to preprocessing a certain amount of data for the machine training model before the design fair. And very likely the scope of the project will shrink from distinguishing two diseases to classifying between healthy control samples and samples with one disorder.

Shi Hu

The goal of this project is to develop a machine learning software solution that can classify brain diseases based on functional magnetic resonance imaging (fMRI) signals retrieved from public databases. To better provide data to the machine learning model, the resulted system is required to first convert fMRI to brain graphs in the form of adjacency matrices. Then, these matrices are fed into pre-trained machine learning models, which shall generate predictions of diseases for the given brain graphs. In general, my responsibility is to find some effective algorithms, implement them and train them. Then, select the best one for one of the two machine learning modules.

I started with reading materials recommended by the supervisor to understand the basic ideas of brain graph. One vital idea was the null model. Null models are synthetic models used by people to analyze brain graphs, which could be easily generated and maintain some basic characteristics of the brain graph. Therefore, at the first stage, I could use them to generate my own synthetic data to test my algorithms before I could have access to real data.

After that, I did research on algorithms may be suitable for this project. Four different algorithms came into my scope. MLP, CNN, DeepWalk and SDNE. MLP (multi-layer perceptron) was a typical basic neural network. CNN (convolutional neural network) was typically used to classify images. As our adjacency matrices could be viewed as mono-channel images, CNN was worth trying. DeepWalk and SDNE were two kinds of graph embedding techniques, which were designed especially for extracting information from graphs. I first tested them on synthetic data. All of them have good performance. Then, after real data were available, I trained and tested them on real data. However, this time only MLP could provide decent results, which was also chosen as our final model.

At this point, my current results support that MLP would be the best choice for our project. However, as suggested by our supervisor, if the data could be well-preprocessed, there may be a different result. Also, it is still expected that some techniques may be appended to MLP to improve its baseline. Additionally, our supervisor suggested an optional task that it could be an interesting idea to cluster the data into several sub-categories. Thus, before the design fair started, I may still work on some of these things.

Jixiong Deng

The goal of our project is to develop a machine learning solution to classify brain diseases using functional magnetic resonance imaging (fMRI) signals retrieved from public sources. The project is divided into two parts, one is the data science module, which transforms fMRIs into adjacency matrices, and the other one is the machine learning module, which is trained and evaluated with the output from the data science module. In our project, I am mainly responsible for developing machine learning algorithms in order to predict disease type by given brain graph data.

Machine Learning is widely used as the automation method and covers lots of various learning mechanisms. To achieve our goal, I need to explore learning algorithms and choose the most suitable models with optimal hyperparameters. I followed instructions from our supervisor. And I found I could not avoid implementing a python project with most popular library in this field, Tensorflow after my inspection in order to establish a strong solution. So it is necessary to make all environment programs and configurations are well set on my local working platform.

When all background situations are set down, I started to implement a supervised learning algorithm named Support Vector Machine (SVM), which can produce a binary decision boundary for multidimensional linear classifier. After long time debugging process, SVM could have an average predictive accuracy around 90% on binary test dataset with well-tuned hyperparameters. My next goal is to optimize data processing part in SVM in order to reduce overfitting phenomenon and increase the predictive accuracy. Because of the good result of our current progress, even higher than our objective, and the delay of our brain graph dataset, we decided not to implement Generative Adversarial Networks.

Besides that, I also attempted to implement a probabilistic reasoning solution named Expectation-Maximization algorithm. By the time of the report, the algorithm is on debugging process and almost down with the prior of bernoulli distribution. Once the current algorithm finished, I will try to implement an advanced version by replacing the prior of bernoulli distribution with the prior of high dimensional Gaussian distribution. The idea is called Gaussian Mixture Models (GMM) and I believe when the more data feed in GMM, the predictive accuracy will be better than our current solution.

Acknowledgments

“Stars, bright night sky; many people, wisdom wide.”

On the very outset of this report, we would like to extend our sincere & heartfelt obligation towards all the personages who have helped us in this endeavour. Without their active guidance, help, cooperation & encouragement, we would not have made headway in the project.

We would like to express our deepest appreciations to our project supervisor, Professor Ashish Khisti. His exemplary guidance, constant encouragement, and careful monitoring throughout our research process provides us with a brilliant pathway and leads us successfully completing our project.

We would like to sincerely thank our project administrator, Dr. Ross Gillett, who always generously guides us to achieve our goals, helps us when we met difficulties and educates us to be great engineers in our future careers, with our most profound gratitude.

We gratefully acknowledge the ADHD-200 Consortium and Autism Brain Imaging Data Exchange (ABIDE) for generously sharing their data with the scientific community.

In addition, we gratefully acknowledge all faculties in University of Toronto for giving us a great opportunity and supports to embark our project with such an amazing experience.

Table of Content

1. Introduction	page 1
1.1 Background and Motivation	page 1
1.2 Project Goal	page 2
1.3 Project Requirements	page 2
2. Final Design	page 4
2.1 System-level Overview	page 4
2.2 Module-level Descriptions	page 6
2.3 Assessment of Final Solution	page 9
3. Testing and Verification	page 11
3.1 Verification table or Validation matrix	page 11
3.2 Final test results	page 12
4. Summary and Conclusions	page 17
5. Reference	page 19
6. Appendices	page 23