# Novel Semi-Supervised Learning Models to Balance Data Inclusivity and Usability In Healthcare Applications

by

Nathan Gaw

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2019 by the
Graduate Supervisory Committee:

Jing Li, Chair
Teresa Wu
Hao Yan
Leland Hu

ARIZONA STATE UNIVERSITY

August 2019

ABSTRACT

Semi-supervised learning (SSL) is sub-field of statistical machine learning that is useful for problems that involve having only a few labeled instances with predictor ($X$) and target ($Y$) information, and abundance of unlabeled instances that only have predictor ($X$) information. SSL harnesses the target information available in the limited labeled data, as well as the information in the abundant unlabeled data to build strong predictive models. However, not all the included information is useful. For example, some features may correspond to noise and including them will hurt the predictive model performance. Additionally, some instances may not be as relevant to model building and their inclusion will increase training time and potentially hurt the model performance. The objective of this research is to develop novel SSL models to balance data inclusivity and usability. My dissertation research focuses on applications of SSL in healthcare, driven by problems in brain cancer radiomics, migraine imaging, and Parkinson's Disease telemonitoring.

The first topic introduces an integration of machine learning (ML) and a mechanistic model (PI) to develop an SSL model applied to predicting cell density of glioblastoma brain cancer using multi-parametric medical images. The proposed ML-PI hybrid model integrates imaging information from unbiopsied regions of the brain as well as underlying biological knowledge from the mechanistic model to predict spatial tumor density in the brain.

The second topic develops a multi-modality imaging-based diagnostic decision support system (MMI-DDS). MMI-DDS consists of modality-wise principal components analysis to incorporate imaging features at different aggregation levels (e.g., voxel-wise,

connectivity-based, etc.), a constrained particle swarm optimization (cPSO) feature selection algorithm, and a clinical utility engine that utilizes inverse operators on chosen principal components for white-box classification models.

The final topic develops a new SSL regression model with integrated feature and instance selection called s2SSL (with "s2" referring to selection in two different ways: feature and instance). s2SSL integrates cPSO feature selection and graph-based instance selection to simultaneously choose the optimal features and instances and build accurate models for continuous prediction. s2SSL was applied to smartphone-based telemonitoring of Parkinson's Disease patients.

I dedicate this dissertation to my parents.

Thank you for loving me so well for all these years.

ACKNOWLEDGMENTS

I would like to thank my advisor, Jing Li, for being a great mentor, teacher, advocate, and encourager. I have learned so much from her, and am grateful that I have had the opportunity to learn how to do top-quality research. I would also like to thank Teresa Wu for her wise insights and advice on my research and future career path. I also express my gratitude to the other members of my committee, Hao Yan and Leland Hu for their input and valuable interactions in my research projects.

Additionally, I am grateful for my collaborators at Mayo Clinic, Kristin Swanson, Todd Schwedt, Catherine Chong, and Andrea Hawkins-Daarud, who have been invaluable to me in providing great opportunities to work together and solve important problems in healthcare. I would also like to thank all of the members of the ASU-Mayo Center for Innovative Imaging (AMCII) for their support and camaraderie—Hyunsoo Yoon, Bing Si, Yinlin Fu, Kun Wang, Shuluo Ning, Lujia Wang, Hope Lancaster, Yanzhe Xu, Fei Gao, Xiaonan Liu, Rashik Kotwal, Suryadipto Sarkar, Jiajing Huang, Jorge Caviedes, Zhiyang Zheng, Na Zou, Can Cui, Min Zhang, and Congzhe Su.

I am also very grateful for the friends I have made during the Ph.D. program, including Sangdi Lin, Logan Mathesen, Gita Ayu, Ghazal Shams, and Daniel Tran, and many others who have been such a great source of support.

TABLE OF CONTENTS

Page

LIST OF TABLES ..... vii

LIST OF FIGURES ..... ix

CHAPTER

1. INTRODUCTION

Background ..... 1

State of the Art ..... 3

Expected Original Contribution ..... 8

Dissertation Organization ..... 11

2. INTEGRATION OF MACHINE LEARNING AND MECHANISTIC MODELS ACCURATELY PREDICTS VARIATION IN CELL DENSITYOF GLIOBLASTOMA USING MULTIPARAMETRIC MRI

Background ..... 12

Literature Review ..... 16

Development of ML-PI ..... 21

Application of ML-PI to Glioblastoma Patient Cohort ..... 31

Conclusion ..... 38

3. A CLINICAL DECISION SUPPORT SYSTEM USING MULTI-MODALITY IMAGING DATA FOR DISEASE DIAGNOSIS

Background ..... 40

Literature Review ..... 46

CHAPTER Page

3. A CLINICAL DECISION SUPPORT SYSTEM USING MULTI-MODALITY IMAGING DATA FOR DISEASE DIAGNOSIS (Continued)

Development of MMI-DDS ..... 49

Clinical Application: A Glioblastoma Study ..... 60

Clinical Application: A Migraine Study ..... 69

Conclusion ..... 78

4. INTEGRATED FEATURE AND INSTANCE SELECTION IN SEMI-SUPERVISED REGRESSION FOR SMARTPHONE-BASED TELEMONITORING OF PARKINSON'S DISEASE PATIENTS

Background ..... 79

Literature Review ..... 82

Methodological Development ..... 91

Simulation Tests ..... 99

Application to Parkinson's Disease Telemonitoring ..... 108

Conclusion ..... 121

5. CONCLUSIONS AND FUTURE WORK ..... 123

REFERENCES ..... 125

APPENDIX

A. SUPPLEMENTARY MATERIALS FOR CHAPTER 2 ..... 142

B. SUPPLEMENTARY MATERIALS FOR CHAPTER 4 ..... 151

LIST OF TABLES

Table Page

1. Prediction Accuracy of ML-PI: Patient-Specific and Uniform Tuning ..... 32

2. Prediction Accuracy of ML-PI with Partially-Uniform Tuning ..... 34

3. Prediction Accuracy of ML-PI, PI, and ML on (a) All Samples and (b) BAT Samples ..... 34

4. PCs Selected by SFFS to Include in the Radiogenomic Models from Advanced and Conventional MRI Sequences ..... 65

5. LOOCV AUCs on All Samples, BAT Samples, and ENH Samples for Conventional And Advanced MRI Sequences ..... 66

6. CV Classification Errors of the Proposed MMI-DDS Applied to MRI Alone, fMRI Alone, and MRI + fMRI Combined ..... 71

7. Imaging Features That Rank in the Top 5% in Terms of the Magnitudes of Contribution Weights for LSVM ..... 74

8. Comparison of the LapRLS and RLS Base Models ..... 101

9. Comparison of LapRLS with cPSO Feature Selection to LapRLS without Feature Selection. ..... 103

10. Comparison of LapRLS + NNGR at Different Levels of $\lambda$. ..... 105

11. Summary of the Performance of s2SSL versus the LapRLS Baseline ..... 107

Table Page

12. Summary of (Sample +) Train + Test Times for s2SSL versus LapRLS ..... 108

13. The Resulting Mean Absolute Error (MAE ± Standard Error) and Standard Error Scree (SES) for Several Different Maximum Feature Settings of s2SSL Applied To (a) Tapping Features, (b) Voice Features, and (c) Tapping + Voice Features. ..... 115-116

14. Summary of the Final Models Chosen by s2SSL for Tapping Features, Voice Features, and Tapping + Voice Features ..... 117

15. Features Chosen for s2SSL Training on (a) Tapping Features, (b) Voice Features And (c) Tapping + Voice Features ..... 119

16. Description of Features Chosen for (a) Tapping, (b) Voice, and (c) Tapping + Voice ..... 120-121

LIST OF FIGURES

Figure Page

1. Workflow of Building ML-PI and Using the Model to Generate a Predicted Cell Density Map for the T2W ROI of Each Tumor/Patient. ..... 28

2. Predicted Cell Density Maps of Selected Patients and Scatter Plots of ML-PI, PI, and ML Predictions ..... 36

3. Contributions of PI and MRI Sequences to ML-PI Cell Density Prediction ..... 38

4. Layout of MMI-DDS ..... 49

5. ROIs Corresponding to the Area Features in Table 7 Shown on an Average Inflated Brain Surface. ..... 75

6. Resting-State Functional Connectivities Corresponding to Table 7 ..... 76

7. The Framework of s2SSL. ..... 99

8. S Curve Used in Model Training to Compare Utility of Semi-Supervised Over Supervised Learning ..... 100

9. Graph Generated by LapRLS Such That Each Instance Has at Least $k_{min} = 6$ Nearest Neighbors ..... 101

10. The Deterioration of the Graph Between Different Instances as More Noise Features Were Added. ..... 103

Figure Page

11. S Curve Used in Model Training to Compare Utility of Sampling over Not Sampling ..... 104

12. Graph Made on 2000 Instances (with $k_{min} = 6$) Before Sampling ..... 106

13. Graphs Made on Sampled Instances Post Sampling at Different Levels of $\lambda$ ..... 106

14. The Elbow Plots of the Mean Absolute Error (MAE ± Standard Error) for (a) Tapping Features, (b) Voice Features, and (c) Tapping + Voice Features ..... 114

15. Scatter Plots of Predicted MDS-UPDRS Score versus True MDS-UPDRS Score For (a) Tapping, (b) Voice, and (c) Tapping + Voice Features ..... 117

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

The advances of sensing and computer technologies have produced immense amounts of data in healthcare. Some data are easier to obtain than others are. For example, in the application of glioblastoma brain cancer, when a doctor wants to get a better idea of the tumor environment, he/she can take biopsy samples and collect medical images of the patient. Biopsies are invasive and the ability to sample is limited, while imaging data is non-invasive and available in large quantity. As another example, in monitoring the progression of Parkinson's Disease (PD), one can use telemonitoring signals collected by the patients' smartphones (e.g., voice, tapping, gait) and clinical instruments such as the Unified Parkinson's Disease Rating Scale (UPDRS). Clinical instruments must be administered in specialized clinics so the data is limited, while telemonitoring by smartphones is convenient and therefore the data can be available in large quantity.

In both of the two aforementioned examples, an important task in building a predictive model is to use the easy-to-get data to predict the hard-to-get data with purpose of minimizing the need for the hard-to-get data in the future. In the brain cancer example, if a predictive model can be built to link image features with biopsy-based histopathological biomarkers, one can use the model to predict the biomarkers everywhere within the tumor using imaging data without the need for additional biopsy samples. This predicted biomarker map will help guide surgical resection and precision radiation therapy. In the PD example, if a predictive model can be built to link telemonitoring signals with

UPDRS scores, one can use the model to predict the UPDRS anytime without requiring the patient's physical presence in a specialized clinic. As UPDRS reflects the severity of PD, using the predicted UPDRS that can frequently be obtained will help the doctors closely monitor disease progression and make timely medical decisions for treatment adjustment.

In building the predictive models, one option is to use a training dataset that only includes labeled data (e.g., biomarkers, UPDRS). Such a dataset will be limited in the sample size due to the difficulty in obtaining the hard-to-get data. Another option is to use both labeled and unlabeled data. The latter refers to samples, for example, with only imaging data available but no biomarker information (another example would be samples with only telemonitoring data available but no UPDRS information). This option is studied in a subfield of machine learning (ML) called semi-supervised learning (SSL).

SSL aims to utilize all the available data (labeled and unlabeled), so it is inclusive in nature. However, not all the included data is useful. For example, some features may correspond to noise and including them will hurt the predictive model performance. Also, although unlabeled samples are relatively easier to obtain and therefore come in large quantity, some samples may correspond to noise and including them in an SSL model will hurt the performance. **The objective of my dissertation** is to develop novel SSL models to balance the data inclusivity and usability, driven by real-world healthcare applications especially in radiomics of brain cancer, migraine imaging, and telemonitoring of PD.

## 1.2 State of the Art

Several subfields of ML are relevant to my proposed work, which are reviewed as follows:

**(1) Semi-supervised learning (SSL)** has been widely used in applications in which labeled data are scarce but unlabeled data are available in large quantity. There are many types of SSL algorithms, including generative, self-training, co-training, low-density separation, and graph-based models. Graph-based SSL has recently become popular because of its relatively high accuracy and efficiency. The basic idea is to construct a graph with vertices being labeled and unlabeled samples in a training set and edges weighted by vertex proximity in the feature space. There are two types of graph-based SSL: transductive and inductive learning models. The former aims to formulate a method to propagate information from labeled samples to unlabeled samples in a specific dataset. In this way, the unlabeled samples in the dataset are classified/predicted. The latter aims to train a model using labeled and unlabeled samples, which is not only used to predict the unlabeled samples in training but also new samples. As examples, for transductive learning, Zhu et al. 2003 proposed a Gaussian random field model with the mean of the field characterized in terms of harmonic functions. They tested the model on digit and text classification tasks. For inductive learning, Belkin et al. 2006 proposed the manifold regularization (MR) framework, which relies on properties of reproducing kernel Hilbert spaces (RKHS) to enable efficient and accurate prediction.

**(2) Feature selection** obtains a non-redundant, relevant subset of features from a set of many features to improve model accuracy and interpretability. Unlike feature

extraction, feature selection does not create new features from the original data. For example, principal components analysis (PCA) is a feature extraction method that generates new features that are a linear combination of original features (Hotelling 1933). Instead, feature selection chooses a subset of existing features that can describe the model (Saeys et al. 2007), providing better model interpretability and improved accuracy.

There are three different types of feature selection, namely (1) filter method, (2) wrapper method, and (3) embedded method. Filter method determines the relevance of different features by examining the inherent properties of the data (e.g., information gain, correlation-based feature selection). Wrapper method includes the model hypothesis in determining the relevance of a specific feature (e.g., the classification accuracy of a discriminant classifier). Embedded method "embeds" the search for an optimal subset within the model construction (e.g., branch-and-bound method) (Guyon and Elisseeff 2003). However, most feature selection algorithms were developed as a pre-processing step or to be integrated with supervised learning. Limited research has been done for feature selection in SSL.

**(3) Sample reduction for graph-based SSL** has been shown to be useful for improving accuracy and efficiency of the SSL algorithms. Some graph-based SSL techniques have limitations on larger datasets due to the computational complexity of incorporating a matrix-embedded graph into model training. Several sampling techniques have been developed to minimize computational time by training on the most relevant samples/instances.

One category of sampling techniques is embedded directly into the objective

function to be minimized. Studies by Zhang et al. 2014 and Zuo et al. 2015 incorporated the $L_1$-norm with respect to the kernel coefficients of the response for manifold regularization (MR). Studies by Lu et al. 2015 and Lu and Wang 2015 employed a Laplacian $L_1$-norm into the objective function.

Other sampling methods are used to reduce the population to a representative set before SSL model training. Performing sampling before model training can have a significant time advantage since the initial graph size before model training is much smaller. Wang and Zhang 2008 used a graph-based sampling method to eliminate bridge points, i.e., instances that are noisy or do not have many nearest neighbors in a given search radius. Goldberg et al. 2009 introduced a cover sampling technique that produces an approximate cover of the unlabeled instances across the manifold. Sun et al. 2014 employed a method that favors sampling instances that have a higher degree in the underlying graph.

In my dissertation, I will address the gaps in the existing research in the following aspects:

- **Lack of integrating underlying medical knowledge of biological processes with semi-supervised models in healthcare:** It is paramount to have tools that can accurately predict spatially heterogeneous biological processes being monitored by medical imaging. There is a lack of machine learning models that combine image-localized biopsies and imaging data with mechanistic models that convey the scientific knowledge of the underlying cellular mechanisms. Without the marriage

of empirical information and scientific knowledge, models are either prone to the variability of empirical information or the smoothness of mechanistic models.

- **Need for methods that select a feature subset to produce a near-global optimal solution in a semi-supervised learning setting:** Given the high-dimensionality of the joint feature set produced by multi-modality data, searching for the subset of features with the near-global optimal classification accuracy is very challenging. An exhaustive search is practically impossible. Greedy search based methods such as sequential forward selection and sequential backward selection suffer from a variety of problems such as stagnation in local optima and a high computational cost. Lately, evolutionary computation (EC) techniques such as genetic algorithms (GA) (Fraser and Burnell 1970), genetic programming (GP) (Koza 1990), differential evolution (DE) (Storn and Price 1997), and neuroevolution (Floreano et al. 2008) have attracted great attention with some initial success in feature selection and classification for medical applications. A new emerging field in EC is swarm intelligence (Bonyadi and Michalewicz 2017) which models the collective behavior of social swarms in nature, such as ant colonies, honeybees, and bird flocks. Although individuals in a swarm are relatively unsophisticated with limited capabilities on their own, they interact together with certain behavioral patterns to cooperatively achieve tasks necessary for their survival. This "intelligent" behavior of the swarm has inspired new algorithmic developments in solving large complex optimization problems with a wide range of application domains such as machine learning (Das et al. 2009), bioinformatics (Das et al. 2008), dynamical systems and

operations research (Parsopoulos 2010). Particle swarm optimization (PSO) is a computational algorithm based on swarm intelligence that mimics the behavior of flying birds and their means of information exchange to solve optimization problems. Each potential solution is seen as a particle with a certain velocity, and flies through the problem space. Each particle adjusts its flight according to its own flying experience and its companions' flying experiences. The particle swarms find optimal regions over complex search spaces through the interaction of individuals in a population of particles. PSO has been successfully applied to a number of difficult combinatorial optimization problems (Jarboui et al. 2008, Chu et al. 2012). PSO has also been shown to be computationally less expensive, converge more quickly, and find better solutions than classic EC algorithms such as GA (Wang et al. 2007, Jarboui et al. 2007).

- **Need to integrate feature selection and sample reduction methods in an SSL regression framework:** Most of the existing SSL models target categorical response variables, i.e., they are in parallel with classification models in supervised learning. Less work has been done with numerical response variables, i.e., in parallel with regression models in supervised learning. Many healthcare applications have numerical response variables such as tumor cell density in glioblastoma and UPDRS in PD. Moreover, this is little work on integrating feature selection and sample reduction with regression-type SSL to balance data inclusivity (SSL) and usability (from two angles, samples and features).

## 1.3 Expected Original Contribution

The objective of my dissertation research is to develop new semi-supervised learning methods that overcome the aforementioned limitations of the existing methods and demonstrate their utility in healthcare applications. The expected original contributions are as follows:

- **Integration of mechanistic models and ML to develop a new SSL model, which was applied to predict intra-tumor cell density of glioblastoma using multiparametric MRI.** In my first topic, I develop a novel semi-supervised learning technique to incorporate labeled and unlabeled data, as well as a bio-based mechanistic model, for prediction of brain tumor content in glioblastoma patients. In medical imaging there have been several advancements that have improved the quality of established imaging techniques and introduced new ones. There is an abundance of heterogeneous imaging types that can be used to discern different properties of the organ of interest. T1+C detects blood brain barrier disruption; T2W measures water content; rCBV detects microvascular volume; EPI+C detects cell density/size and microvessel volume; MD detects bulk water movement; and FA detects directionality of water movement. In addition, texture algorithms can also be used to process the available images to infer additional information. Gray level co-occurrence matrix (GLCM) measures how often pairs of pixels with specific values within a specific window of the image occur; local binary patterns (LBP) measure local spatial patterns and are robust to monotonic gray-scale changes in the image; and Gabor filters determine if there is a frequency content in

a specific direction for a particular region of interest. The fundamental meaning of these texture features for particular applications are more abstract and research is still being performed to improve their interpretation. Imaging information can be used to infer information about the underlying phenotypic information expressed by different micro-expressions detected in the tissue. This area of research is known as radiomics. For example, in glioblastoma, it has been found that imaging features can be linked with the tumor density and genetic information inferred from a biopsy sample (Hu et al. 2015, Hu et al. 2016). One limitation in radiomics, however, is having an insufficient number of biopsy samples to train an accurate machine learning model. Due to the invasive nature of biopsies, they can be very expensive to collect at the cost of patient safety. Thus, there is a need to utilize additional sources of information to improve model prediction. Imaging information of unbiopsied samples can be utilized by a semi-supervised learning approach. Additionally, there is already scientific knowledge of the cell diffusion and proliferation patterns of glioblastoma available through mechanistic models. These models are derived from imaging information of the patient and utilize knowledge of glioblastoma to make its prediction. Information from these mechanistic models was utilized as a sort of prior knowledge to guide the prediction of the machine learning model.

- **Development of a new constrained particle swarm optimization (cPSO) based feature selection algorithm, which was applied to two healthcare applications in a supervised learning setting and prepared to be further extended to an SSL**

**setting.** In my second work, I address the issue of data usability from the angle of selecting informative features and propose a near-globally optimal feature selection method called constrained particle swarm optimization (cPSO). I develop cPSO as part of a proposed multi-modality imaging-based diagnostic decision support system (MMI-DDS). The cPSO algorithm honors a pre-specified maximum number of features to avoid model overfitting, while also evaluating feature quality and not including poor quality features in the trained model (thus ensuring that the number of features selected will always be less than or equal to the specified number). The algorithm is applied to two healthcare applications: (1) predicting genetic mutation in biopsies of glioblastoma patients and making accurate predictive maps of genetic mutations in the tumor and peripheral areas, and (2) classifying migraine patients and determining potential clinical indicators of the disease.

- **Development of a new SSL model with cPSO-integrated feature and instance selection, which was applied to smartphone-based telemonitoring of Parkinson's Disease patients.** I introduce a first-of-its-kind semi-supervised feature and instance selection algorithm for regression tasks that combines SSL, cPSO, and graph-sampling. Because the proposed model integrates both feature and sample selection with SSL, it is named as s2SSL, implying being an SSL with selection in two aspects: feature and sample. s2SSL aims to balance data inclusivity (through the SSL) and usability (through the feature and sample selection). This work is done in the context of smartphone-based telemonitoring, an emerging

healthcare area that has high potential to closely monitor a patient's disease severity—in particular, Parkinson's Disease (PD). Because smartphones have improved in technological capabilities (e.g., better quality accelerometers, cameras, microphones, etc.), they can now be used to remotely monitor patients, reducing the frequency of times a patient needs to visit the clinic to receive assessment on disease severity.

## 1.4 Dissertation Organization

The proposed dissertation research will be presented in three chapters: Chapters 2, 3, and 4 encapsulate the three topics of my research (described in the previous section). Chapter 2 integrates a biologically-based mechanistic model with semi-supervised learning, Chapter 3 develops a feature selection approach to select a near-globally optimal feature subset from an abundance of heterogeneous imaging features, and Chapter 4 integrates a cPSO-based feature selection and a graph-sampling approach in an SSL framework.

# CHAPTER 2

# INTEGRATION OF MACHINE LEARNING AND MECHANISTIC MODELS ACCURATELY PREDICTS VARIATION IN CELL DENSITY OF GLIOBLASTOMA USING MULTIPARAMETRIC MRI

## 2.1 Background

Glioblastoma (GBM) ranks among the most lethal of all human cancers. The median survival in the general patient population with first-line treatment is 14 months, with a 26% 2-year survival rate (Stupp et al. 2005, Sottoriva et al. 2013). Poor survival and treatment failure can largely be attributed to tumor invasion and intratumoral heterogeneity (Inda et al. 2014). Intratumoral heterogeneity manifests as the spatial heterogeneity in tumor cell density in and around the tumor regions visible on clinical imaging as well as the different molecular signatures of tumor cells within different regions of the same tumor. As a result, different sub-regions of a tumor may have different therapeutic sensitivities, leading to treatment failure and poor survival. Success of the Precision Medicine (PM) revolution hinges on the ability to address such heterogeneity within and between patients (Martelotto et al. 2014, Brocks et al. 2014, Baldock et al. 2013, Jackson et al. 2015).

To capture intratumoral heterogeneity, a critical first step is to obtain tumor-rich biospecimens for histological and molecular analysis, which has been a challenging task in the current clinical practice. For example, in the NIH-funded large-scale cancer genomics project, the Cancer Genome Atlas (TCGA), only 35% of the initially submitted biopsy samples contained adequate tumor content to make genetic analysis possible (McLendon et al. 2008). Ideally, due to the invasive nature of biopsies, and since the abnormality seen

on clinical imaging reveals only the tip of the iceberg of the overall tumor invasion (Swanson et al. 2000, Swanson et al. 2002, Swanson et al. 2003, Baldock et al. 2014, Corwin et al. 2013 Wang et al. 2009, Sodt et al. 2010), one would want to map out the tumor cell density distribution across the clinical imaging (magnetic resonance imaging, MRI) such that biopsy samples can be prioritized. Such a tumor cell density map would offer two additional important clinical benefits: it will assist with enhancing precision of surgical resection and optimizing the dose distribution of radiotherapy.

In GBM, various MRI sequences containing complementary information have been used to assist clinical decision making, including the conventional T1- and T2-weighted imaging and more advanced imaging such as diffusion tensor imaging (DTI), which measures white matter infiltration, and perfusion imaging, which measures microvessel morphology. Mapping intratumoral cell density distribution can take advantage of multi-sequence or multiparametric MRI. There have been two parallel types of efforts taking as inputs multiparametric MRI to generate tumor cell density maps – machine learning and mechanistic modeling.

Machine learning (ML) models can be trained to link localized imaging features of multiparametric MRI at each biopsy location with pathologist quantified tumor cell density (Durst et al. 2014, Hu et al. 2015, Chang et al. 2017). This results in a predictive tumor cell density ML model map that can be applied over the entire tumor. Since ML models are trained on the scant data provided by image-localized biopsies from different regions of previous tumors, they are prone to vulnerability with regard to any biases or imbalance in the data feeding the model. Based on the breadth and depth of these training data, the resultant trained ML model can be used to predict the cell density of any location, including

locations that are not biopsied.

Mechanistic models, on the other hand, are built on first principles understanding of cancer biology that constrain interpretation as to how the multiparametric MRIs might provide insights into the tumor cell density across the brain. One well-known mechanistic model is the Proliferation-Invasion (PI) model (Swanson et al. 2000, Swanson et al. 2002, Swanson et al. 2003, Baldock et al. 2014, Corwin et al. 2013 Wang et al. 2009, Sodt et al. 2010). The PI model is based on the principle that gliomas are proliferative and invasive, and thus simulations of the PI model are based on patient-specific estimates of the tumor cell net proliferation and invasion rates, estimated for each patient using the contrast enhanced T1-weighted and T2-weighted MRIs. Based on the premise underlying the PI model, given outlines of imaging abnormality on these pretreatment images along with gray/white matter segmentation of the patients' brain, the PI model can produce a tumor cell density map anywhere within the patients' brain (Baldock et al. 2014, Wang et al. 2009, Sodt et al. 2010, Swanson et al. 2013).

Both ML and mechanistic models have strengths and limitations. ML is a data-driven approach, which has the strength of utilizing the available data, but is limited in that a model built on a particular dataset may not generalize well to other datasets. For instance, ML models for tumor cell density can make predictions that are counter to biological intuition and experience including suggesting unrealistic fluctuations in cell density over small distances or predicting biologically unlikely significant regions of high tumor cell density distant from the imageable component of the tumor. On the other hand, the PI model has better generalizability because it is a mechanistic model based on the underlying principles of cancer biology. But the PI model is limited in that it assumes cell density

monotonically decreases from around the center of the tumor mass (i.e., enhancing core on a T1+C image) to the surrounding non-enhancing parenchyma, so called brain around the tumor (BAT), not allowing significant local fluctuations. While it is generally true that higher cell densities are in the center of the imaging abnormality and the lower cell densities are on the outskirts, the monotonic nature limits the high resolution accuracy of the PI model estimates in BAT. Here I propose to integrate these ML and PI approaches into a hybrid model to leverage the strengths of each model and overcome the limitations of using each model alone. To the best of my knowledge, such a hybrid model does not exist, which motivates this research.

The focus of this chapter is to develop a novel hybrid model, called ML-PI, that integrates ML and PI models to increase accuracy in predicting intratumoral cell density distribution for each patient. ML-PI adopts a semi-supervised learning (SSL) framework, which utilizes both biopsy samples (called labeled data) and biopsy-free sub-regions of the tumor (called unlabeled data). SSL has been widely used in various applications in which labeled data are scarce but unlabeled data are readily available and in a large quantity. This is also the case for my application in which biopsy samples are very limited for each patient and there are abundant sub-regions of the tumor that are not biopsied but with image features readily available. The contributions of this research are summarized as follows:

- Under a graph-based SSL framework, ML-PI incorporates PI-estimated cell density to regularize the multiparametric MRI based SSL model. ML-PI is able to learn patient-specific predictive relationships between imaging features and cell density that is superior to each modeling method alone. The resultant ML-PI model

improves the ability to capture substantial intra- and inter-patient heterogeneity.

- I propose an algorithm called Relief-ML-PI, adapted from the Relief algorithm (Robnik-Sikonja and Kononekno 2003), to quantify the contribution from each MRI sequence and PI to the final cell density prediction. One of the major distinctions of Relief-ML-PI from the original Relief, is that it is used to examine feature contributions of the model post-training, as opposed to being used for feature selection pre-model training. Finding their respective contributions to prediction of tumor cell density helps knowledge discovery about GBM. Also, knowing the contribution from PI relative to imaging features reveals the importance of incorporating mechanistic models into data-driven ML.
- I apply ML-PI to a clinical cohort of primary GBM patients undergoing surgical biopsy and resection. High accuracy in cell density prediction is achieved in comparison with competing methods. Using Relief-ML-PI, PI is found to contribute most significantly to the prediction. Predicted cell density maps are generated for each patient across the tumor mass and BAT, allowing for precision treatment.

## 2.2 Literature Review

This research intersects with three existing research areas: 1) ML models that use multiparametric MRI to predict intratumoral regional cell density; 2) mechanistic tumor proliferation and invasion models that achieve the same purpose as 1); 3) SSL used to incorporate unlabeled samples to improve the model performance. In what follows, I will

review the existing work in each area and point out limitations.

### 2.2.1 ML models for Intratumoral Regional Cell Density Prediction Using Multiparametric MRI

To the best of my knowledge, this area has only limited works so far. One related work (Hu et al. 2015) developed an ML pipeline to predict regional cell density within each tumor. This pipeline included three key steps, including 1) texture feature extraction from co-registered multiparametric MRI images (T1+C, T2W, rCBV, EPI+C, MD, FA) localized at biopsied tumoral sub-regions; 2) feature dimension reduction by modality-wise principal component analysis (PCA); 3) building of a classifier using an ensemble of classification algorithms. This pipeline was used to classify high ($\geq$80%) vs. low ($< 80\%$) density for a cohort of 82 biopsy samples from 18 patients – the same cohort I focus on in this study. Furthermore, other researchers have developed methods to predict regional cell density on a continuous scale (0-100%). For example, Durst et al. used 12 imaging variables followed by PCA and generalized estimating equations regression (GEER) to predict regional cell density of 10 patients (Durst et al. 2014). Chang et al. 2017 used T1+C, T2-FLAIR, and ADC features to predict cell density of 36 patients by a multiple linear regression.

The existing works have limitations: First, except the work in Hu et al. 2015, the other papers reported only training accuracy of the predictive models. It is well-known that training accuracy over-estimates the true accuracy of a predictive model. Additionally, while identification of high tumor regions (e.g., with $\geq$80% density) like the work in Hu

et al. 2015 has significant clinical value for guiding neurosurgery and biopsy, prediction of cell density on a continuous scale is more challenging and clinically valuable for guiding more nuanced decision making in surgery and dose optimization in radiotherapy. Also, all the existing studies were based on biopsy samples alone, without incorporating 1) abundant regional samples with imaging feature readily available but not biopsied (i.e., unlabeled data), and 2) first principles of tumor cell biology as characterized by mechanistic models of proliferation and invasion.

### 2.2.2 Mechanistic PI Model for Patient-Specific Tumor Cell Density Estimation

The proliferation-invasion (PI) model aims to capture the most basic understanding of what cancer is: cells that grow uncontrollably and invade surrounding tissue. The invasion term is particularly relevant here as glioblastomas are known to be diffusely invasive with the potential to migrate long distances in the human brain (Baldock et al. 2013, Jackson et al. 2015). Mathematically, the PI model is written as follows:

$$\overbrace{\frac{\partial c}{\partial t}}^{\text{Rate of Change of Cell Density}} = \overbrace{\nabla \cdot (D(x)\nabla c)}^{\text{Invasion of Cells into Nearby Tissue}} + \overbrace{\rho c\left(1-\frac{c}{K}\right)}^{\text{Proliferation of cells}}$$

where $c(x,t)$ is the tumor cell density, $D(x)$ is the net rate of diffusion taken to be piecewise constant with different values in gray and white matter, $\rho$ is the net rate of proliferation and $K$ is the cell carrying capacity. This model has been used to predict prognosis (Wang et al. 2009), radiation sensitivity (Rockne et al. 2010), benefit from resection (Baldock et al. 2014), and IDH1 mutation status (Baldock et al. 2014b). Additionally, this model was used to create untreated virtual controls for use in defining

response metrics that are more prognostically significant than those currently in use (Neal et al. 2013a, Neal et al. 2013b).

I note that the vast majority of the clinically relevant PI literature focuses on the intuition derived from the patient-specific parameter values ($D$ and $\rho$), i.e. the gross tumor growth profile, rather than a voxel by voxel cell density prediction. This is exactly because the PI model tends to smooth local regional cell density differences on this scale. The use of the PI model cell densities in the hybrid model presented here is for a similar purpose: these predictions provide an insight into the expected overall pattern but need to be augmented by more sophisticated data-driven ML methods to achieve local accuracy. That is, the biological insights provided by the PI model provides a means to regularize the biologically unrealistic spatially heterogeneity seen in the ML models for tumor cell invasion.

### 2.2.4 Semi-Supervised Learning (SSL)

SSL has been widely used in applications in which labeled data are scarce but unlabeled data are available in large quantity. There are many types of SSL algorithms, including generative (Holub et al. 2005, Fujino et al. 2005), self-training (Li et al. 2008, Tanha et al. 2017, Bache and Lichman 2013), co-training (Wan 2009, Zhou et al. 2007), low-density separation (Zhu and Lafferty 2005, Lawrence and Jordan 2005), and graph-based models. This study utilizes a graph-based SSL method to integrate PI with ML, and so a brief summary of different graph-based SSL methods will be provided in this section.

Graph-based SSL has recently become popular because of its relatively high

accuracy and efficiency. The basic idea is to construct a graph with vertices being labeled and unlabeled samples in a training set and edges weighted by vertex proximity in the feature space. There are two types of graph-based SSL: transductive and inductive learning models. The former aims to formulate a method to propagate label information from labeled samples to unlabeled samples in a specific dataset. In this way, the unlabeled samples in the dataset are classified/predicted. The latter aims to train a model using labeled and unlabeled samples, which is not only used to predict the unlabeled samples in training but also new samples.

For transductive learning, Zhu et al. 2003 proposed a Gaussian random field model with the mean of the field characterized in terms of harmonic functions. They tested the model on digit and text classification tasks. Reference Zhou et al. 2004 introduced a local and global consistency framework based on the quadratic loss of prediction on labeled samples regularized by a normalized Laplacian matrix. For inductive learning, Zhu and Lafferty 2005 regularized generative mixture models with graph Laplacian and demonstrated its performance on handwritten digit and teapots image datasets. Belkin et al. 2006 proposed the manifold regularization (MR) framework, which relied on properties of reproducing kernel Hilbert spaces (RKHS) to enable efficient and accurate prediction.

This chapter aims to adopt the SSL concept by leveraging the abundant intratumoral regional samples that are not biopsied (unlabeled data) to compensate for the limited biopsy samples (labeled data). I chose to use the graph-based SSL by Belkin et al. 2006 as my baseline model because of its proven high accuracy and efficiency in various applications, as well as its inductive learning ability that allows the trained model to be used to predict new patients. Furthermore, using the SSL model by Belkin et al. 2006 as baseline, I

innovate it by 1) incorporating the mechanistic PI model, leading to a hybrid ML-PI model; and 2) proposing a post-analysis step of ML-PI for identifying contributions from different MRI sequences and PI.

## 2.3 Development of ML-PI

### 2.3.1 Patient Recruitment

Patients were recruited with clinically suspected GBM undergoing preoperative stereotactic MRI for first-line surgical resection prior to any treatment, as per institutional review board protocol. Approval was obtained from the institutional review boards at Barrow Neurological Institute (BNI) and Mayo Clinic in Arizona (MCA) in accordance with the Declaration of Helsinki. All patients provided written and informed consent prior to enrollment. The patient cohort presented here has also been described in previous studies (Hu et al. 2015). 82 biopsy samples were collected from 18 GBM patients, with each patient having 2-14 biopsy samples.

### 2.3.2 Surgical Biopsy

Pre-operative conventional MRI, including T1-Weighted contrast-enhanced (T1+C) and T2-Weighted sequences (T2W), was used to guide biopsy selection. Each neurosurgeon collected an average of 5–6 tissue specimens from each tumor by using stereotactic surgical localization, following the smallest possible diameter craniotomies to minimize brain shift. Specimens were collected from both enhancing mass (as seen on T1+C) and non-enhancing BAT (as seen on T2W) for each tumor. The neurosurgeons

recorded biopsy locations via screen capture to allow subsequent coregistration with multiparametric MRI datasets. The biopsy tissue specimens were reviewed blinded to diagnosis by a neuropathologist and assessed for tumor content. Taking into account all visible cells (neurons, inflammatory cells, reactive glia, tumor cells, etc.), the percent tumor nuclei were estimated. Additional details of methods for surgical biopsy and pathological density measurement can be found in (Hu et al. 2015).

### 2.3.3 Multiparametric MRI and ROI Segmentation

Six multiparametric images were included in the present study, including T1+C, T2W, EPI+C, MD, FA, and rCBV (detailed MRI protocols and image co-registration can be found in Hu et al. 2015 and the supplementary information). The main goal was to generate cell density predictions for the extent of the abnormality shown on T2W (called T2W ROI hereafter), which includes both the tumor mass enhanced on T1+C and non-enhanced BAT. The latter is known to harbor residual tumor cells after resection, which lead to treatment failure and recurrence (Hu et al. 2015). The T2W ROI of each tumor was manually segmented by a board-certified neuroradiologist.

### 2.3.4 Image Feature Computation and PI Density Estimation.

An 8x8 voxel box was placed at the location of co-registered images that corresponds to each biopsy sample. The average gray-level intensity over the 64 voxels within the box was computed for each image sequence. In addition to computing features

for the biopsy samples (i.e., labeled samples), I also computed features for unlabeled samples in the following way: One slice of MRI was chosen for each patient, which is approximately the cross-section that included a balanced amount of enhancing mass and non-enhancing BAT. Furthermore, 8x8 voxel boxes were placed one pixel apart on the T2W ROI of the chosen slice, and the same image features as those of the biopsy samples were computed for each box.

Using the T1+C and T2W images of each patient as input, voxel-wise density estimation was generated by the PI model. Average PI density over the pixels in each 8x8 box on the selected slice was computed.

### 2.3.5 Data Augmentation by Virtual Biopsies

To provide a balanced dataset for ML-PI model training, virtual biopsies were identified for each patient (if necessary) to balance the high density samples with 'virtual' low density samples according to the steps described in the supplementary information. A total of 39 virtual biopsy samples were added with each patient having 0-6 samples. In Appendix A, Figure A1(a) shows a histogram of pathological density for the real biopsy samples in my dataset, which indicates a clear imbalance toward high density. Figure A1(b) shows a histogram of augmented samples, which indicates good balance. Furthermore, for each virtual biopsy sample, I used the same approach as that for real biopsy samples to compute imaging features and average PI density. Note that virtual biopsy samples were only used in model training, but not in validation of the model performance. The latter was based purely on real biopsy samples.

2.3.6 Development of a Hybrid ML-PI Model

The basic idea of ML-PI is to incorporate PI-estimated regional cell density into a graph-based SSL. ML-PI is a significant expansion from a typical supervised model that takes the following form:

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \quad \frac{1}{L}\sum_{l=1}^{L}(y_l - f(\mathbf{z}_l))^2 + \gamma_A \|f\|_K^2. \tag{2.1}$$

$L$ is the number of biopsy samples in a training dataset. $y_l$ is the pathologically measured tumor cell density for the $l$-th sample. $\mathbf{z}_l$ contains gray-level intensity of each MRI sequence averaged over the 8x8 voxel box placed at the $l$-th biopsy sample location. $f(\mathbf{z}_l)$ is a predictive function for cell density. $(y_l - f(\mathbf{z}_l))^2$ is a loss function that measures the discrepancy between the pathological and predicted density of each biopsy sample. $f$ is a function on the reproducing kernel Hilbert space (RKHS), $\mathcal{H}_K$, with a Mercer kernel $K$. $\|f\|_K^2$ is a norm on $\mathcal{H}_K$, which encourages stability and generalizability of the solution. $\gamma_A$ is a tuning parameter.

Equation (2.1) is a supervised learning model because it uses only the biopsy samples (labeled data). To incorporate unlabeled data and PI-estimated density into the model, I follow the idea of SSL and build a graph on all labeled and unlabeled samples. Specifically, one graph $G = (\mathbf{V}, \mathbf{W})$ is built for each patient. $\mathbf{V}$ is the set of vertices and $\mathbf{W}$ contains the weight of edge between each pair of vertices. Let $n = L + U$ be the number of vertices of the graph. $L$ is the number of all biopsy samples and $U$ is the number of voxels on the T2W ROI for the target patient. The edge weight between vertices $v_i$ and $v_j$, $i, j = 1, \dots, n$, can be computed using a product of two Gaussian functions, i.e.,

$$w_{ij} = w_{ij,z} \times w_{ij,PI} = exp\left(-\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{2\psi_z^2}\right) \times exp\left(-\frac{(PI_i - PI_j)^2}{2\psi_{PI}^2}\right). \quad (2.2)$$

$PI_i$ is PI-estimated cell density averaged over the 8x8 box centered at the $i$-th voxel. $\mathbf{z}_i$ contains gray-level intensity of each MRI sequence averaged over the 8x8 box centered at the $i$-th voxel.

In essence, $w_{ij}$ reflects the closeness between two samples/vertices in terms of their respective image features ($w_{ij,z}$) and PI estimations ($w_{ij,PI}$). $\psi_z$ and $\psi_{PI}$ are parameters to adjust contributions to the weight from image features and PI, respectively.

Furthermore, the graph $G$ can be encoded into a Laplacian matrix defined as $\mathbf{\Omega} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the vertex degree matrix, i.e., a diagonal matrix with diagonal elements being the total sum of edge weights associated with each vertex, and $\mathbf{W}$ is the matrix of all the edge weights. Then, the model in (2.1) can be augmented by incorporating the graph Laplacian matrix, which gives the proposed ML-PI model as:

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \quad \frac{1}{L}\sum_{l=1}^{L}\big(y_l - f(\mathbf{x}_l)\big)^2 + \gamma_A\|f\|_K^2 + \frac{\gamma_I}{\sum_{i,j} w_{ij}}\mathbf{f}^T\mathbf{\Omega}\mathbf{f}. \quad (2.3)$$

$\mathbf{x}_l = (\mathbf{z}_l, PI_l)$. $\mathbf{f}$ contains predictive density for each labeled and unlabeled sample, i.e., $\mathbf{f} = \big(f(\mathbf{x}_1), \dots, f(\mathbf{x}_L), f(\mathbf{x}_{L+1}), \dots, f(\mathbf{x}_{L+U})\big)^T$. $\sum_{i,j} w_{ij}$ is a sum of all the edge weights in the graph. Because of patient heterogeneity, I found that the graph of each patient has a wide range of sparsity levels, which causes difficulty in choosing a common search range for the tuning parameter $\gamma_I$. Adding $\sum_{i,j} w_{ij}$ solves this problem by normalizing patient-specific graphs to allow for $\gamma_I$ to be tuned within a common range.

Through some algebra, the last term in (2.3) can be shown to become:

$$\mathbf{f}^T\mathbf{\Omega}\mathbf{f} = \sum_{i,j=1}^{L+U}(f(\mathbf{x}_i) - f(\mathbf{x}_j))^2 w_{ij,z} \times w_{ij,PI}. \quad (2.4)$$

Then, it is clear that the minimization in (2.3) pushes samples that are closer in image features (i.e., with a larger $w_{ij,z}$) and in PI estimations (i.e., with a larger $w_{ij,PI}$) to have more similar predictions. This is traded off with the loss on the labeled data (the first term in (2.3)) and the smoothness of the predictive function in RKHS (the second term in (2.3)). In the extreme case when $w_{ij,z} = w_{ij,PI} = 0$ for all the edges, (2.3) becomes the supervised learning model in (2.1). In essence, the role of PI in the proposed model is to regularize the learning of the predictive function in order to make sure the spatial proximity of predicted densities conform with that of PI densities to some extent. This implicitly takes into account the bio-mechanism of tumor growth, which is the foundation of the PI model.

The Representer Theorem (Scholkopf et al. 2001) can be used to show that an analytical solution for (2.3) exists in $\mathcal{H}_K$, described in Theorem 1 below (proof is provided in the supplementary information).

**Theorem 2.1:** The solution of the optimization in (2.3) is the following expansion in terms of both labeled and unlabeled samples:

$$f^*(\mathbf{x}) = \sum_{i=1}^{L+U} \alpha_i K(\mathbf{x}_i, \mathbf{x}), \tag{2.5}$$

where $\mathbf{x}$ is any sample for which the cell density is to be predicted, which can be an unlabeled sample included in the ML-PI model in (2.3) or not (e.g., a sample outside the ROI or on a different slice of the tumor). $\alpha_i$'s are coefficients.

With the form of the solution to (2.3) given in Theorem 1, the coefficients, $\alpha_i$'s, need to be estimated. To achieve this, insert (2.5) into (2.3), and obtain the following convex differentiable objective function of $\boldsymbol{\alpha} = [\alpha_1 \cdots \alpha_{L+U}]^T$:

$$\boldsymbol{\alpha}^* = \text{argmin}\ \frac{1}{L}(\mathbf{y} - \mathbf{JK}\boldsymbol{\alpha})^T(\mathbf{y} - \mathbf{JK}\boldsymbol{\alpha}) + \gamma_A \boldsymbol{\alpha}^T \mathbf{K}\boldsymbol{\alpha} + \frac{\gamma_I}{\sum_{i,j} w_{ij}} \boldsymbol{\alpha}^T \mathbf{K}\boldsymbol{\Omega}\mathbf{K}\boldsymbol{\alpha}$$

**J** is an $(L+U) \times (L+U)$ diagonal matrix in which the first $L$ entries are 1 and the rest are 0. **K** is an $(L+U) \times (L+U)$ Gram matrix over labeled and unlabeled samples. **y** is an $(L+U) \times 1$ vector defined by $\mathbf{y} = [y_1 \cdots y_L, 0 \cdots 0]^T$. Furthermore, taking the derivative with respect to $\boldsymbol{\alpha}$, the following is obtained

$$\tfrac{1}{L}(\mathbf{y} - \mathbf{JK}\boldsymbol{\alpha})^T(-\mathbf{JK}) + \left(\gamma_A \mathbf{K} + \tfrac{\gamma_I L}{\sum_{i,j} w_{ij}} \mathbf{K}\boldsymbol{\Omega}\mathbf{K}\right)\boldsymbol{\alpha} = 0.$$

Solving for $\boldsymbol{\alpha}$, the solution is

$$\boldsymbol{\alpha}^* = (\mathbf{JK} + \gamma_A L\mathbf{I} + \tfrac{\gamma_I L}{\sum_{i,j} w_{ij}} \boldsymbol{\Omega}\mathbf{K})^{-1}\mathbf{y}. \quad (2.6)$$

**I** is an $(L+U) \times (L+U)$ identity matrix. Inserting the $\alpha_i$'s obtained above into (4), the predictive function, $f^*(\mathbf{x})$, is finally obtained. $f^*(\mathbf{x})$ can be used to generate a predicted cell density for every voxel within the ROI and thus forming an intratumoral cell density map. The tuning parameters of (2.3)—namely, $\gamma_A$, $\gamma_I$, and $\eta$ (width of the radial basis function kernel $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\eta^2}$)—are then adjusted to find the value for $f^*(\mathbf{x})$ that maximizes accuracy of ML-PI. Figure 1 summarizes the workflow of building the ML-PI model and using the model to generate a predicted cell density map for the T2W ROI of each tumor.

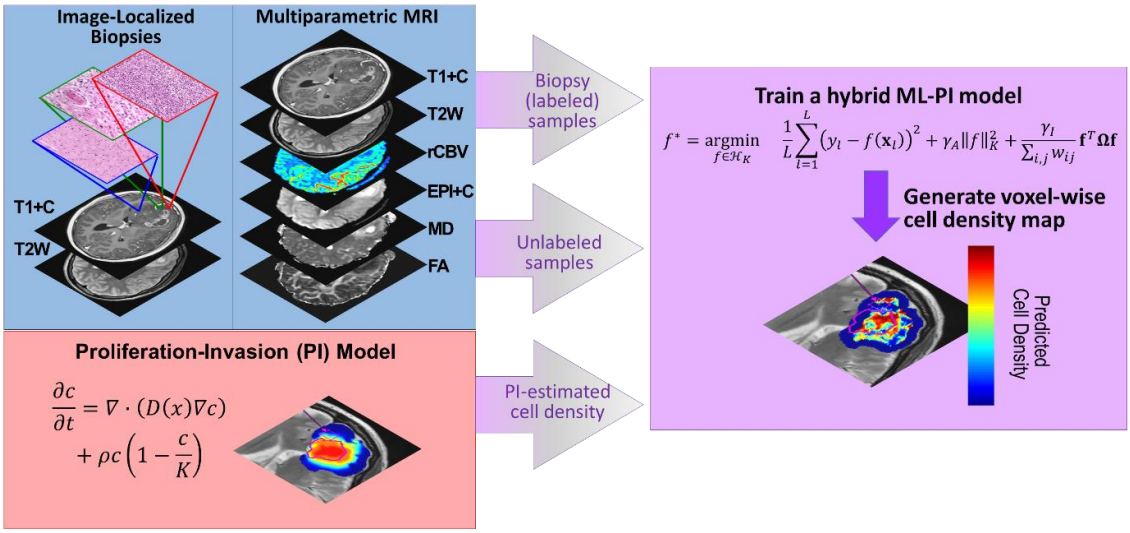



Figure 1: Workflow of building ML-PI and using the model to generate a predicted cell density map for the T2W ROI of each tumor/patient.

### 2.3.7 Feature Contribution Analysis for ML-PI

It is important to determine the quantitative contribution of each feature (i.e., imaging features and PI-estimated density) to the prediction made by ML-PI. It is a reasonable belief that all of the included MRI sequences and PI are biologically relevant to tumor cell density. Therefore, inclusion of all of them as features in building the ML-PI model is valuable, while their relative contributions may vary. Thus, instead of employing feature selection (a step prior to building a predictive model with purpose of removing irrelevant features), I chose to use a post-processing step that identifies how much each feature contributes to the prediction.

Let $x$ be a feature used in ML-PI, which can be a feature computed from an MRI sequence or PI-estimated cell density. The objective is to compute a score for $x$, $s(x)$, that

represents the contribution of $x$. To achieve this, I develop an algorithm based on the well-known Relief algorithm (Robnik-Sikonja and Kononenko 2003), which I call "Relief-ML-PI". Note that Relief was developed as a feature selection algorithm for supervised learning models. My innovation in this chapter is to modify it to become a post-analysis algorithm for feature contribution analysis of SSL models, specifically the ML-PI model. The proposed definition of $s(x)$ is the following: let **T** be the training dataset from which ML-PI is built. **T** includes both labeled and unlabeled samples. Let $i$ and $i_r$ be samples in **T**; $i_r$ is the $r^{\text{th}}$ nearest neighbor of $i$ on the graph $G$. Furthermore, consider the predicted cell density of the two samples by ML-PI, $\hat{y}_i$ and $\hat{y}_{i_r}$, and their respective measurements on feature $x$, $x_i$ and $x_{i_r}$. The definition of $s(x)$ can be based on the difference between two probabilities, i.e.,

$$s(x) = P\left(x_i \text{ and } x_{i_r} are\ different \middle| \hat{y}_i \text{ and } \hat{y}_{i_r}\ are\ different\right) - P\left(x_i \text{ and } x_{i_r} are\ different \middle| \hat{y}_i \text{ and } \hat{y}_{i_r}\ are\ similiar\right). \quad (2.7)$$

The first term represents the probability that feature $x$ is able to separate samples with different prediction values, while the second term represents the probability that $x$ separates samples with similar prediction values. The larger the first probability and the smaller the second, the higher the $s(x)$. Furthermore, using the Bayes' rule, (2.7) can be written as:

$$s(x) = \frac{P(\hat{y}_i \text{ and } \hat{y}_{i_r}\ are\ diff. | x_i \text{ and } x_{i_r} are\ diff.) \times P(x_i \text{ and } x_{i_r} are\ diff.)}{P(\hat{y}_i \text{ and } \hat{y}_{i_r}\ are\ diff.)} - \frac{\{1 - P(\hat{y}_i \text{ and } \hat{y}_{i_r}\ are\ diff. | x_i \text{ and } x_{i_r} are\ diff.)\} \times P(x_i \text{ and } x_{i_r} are\ diff.)}{1 - P(\hat{y}_i \text{ and } \hat{y}_{i_r}\ are\ diff.)}. \quad (2.8)$$

The format of $s(x)$ in (2.8) makes it relatively easier than (2.7) to develop an

algorithm to estimate $s(x)$. The algorithm, Relief-ML-PI, is presented in Algorithm 2.1. The basic idea is to randomly select $m$ samples from $\mathbf{T}$. For each $i = 1, \dots, m$, find its $k$ nearest neighbors $i_r$, $r = 1, \dots, k$. Then, estimate the probabilities in (2.8) and eventually the $s(x)$ using lines 7-9 of the algorithm, in which

$$d(\hat{y}_i, \hat{y}_{i_r}) = \frac{|\hat{y}_i - \hat{y}_{i_r}|}{max(\hat{y}_j | j \in \mathbf{T}) - min(\hat{y}_j | j \in \mathbf{T})},$$

$$d(x_i, x_{i_r}) = \frac{|x_i - x_{i_r}|}{max(x_j | j \in \mathbf{T}) - min(x_j | j \in \mathbf{T})},$$

as the normalized difference between the response variables or feature values of two samples, and

$$\delta(i, i_r) = \frac{\delta'(i, i_r)}{\sum_{l=1}^{k} \delta'(i, i_r)}, \delta'(i, i_r) = e^{-(\frac{rank(i, i_r)}{\sigma})^2}.$$

$\delta'(i, i_r)$ weights each of the $k$ nearest neighbors for sample $i$ and $\delta(i, i_r)$ normalizes the weights. I choose to use the rank of the $k$ nearest neighbors instead of computing the numerical distance due to the same reason as Relief, i.e., to make sure different samples are equally accounted for.

**Algorithm 2.1** Relief-ML-PI

**Input:** measurement data $x_i$ and predicted response $\hat{y}_i$ for each sample in training set **T**; tuning parameters $m$, $k$.

**Output:** $s(x)$

**Initialize:**
$s(x) \leftarrow 0$; $N_{dy}(x) \leftarrow 0$; $N_{dx}(x) \leftarrow 0$; $N_{dy\&dx}(x) \leftarrow 0$;
**for** $i = 1$ **to** $m$ **do**
    Randomly select a sample $i$ from **T**;
    Find $k$ nearest neighbors for sample $i$, $i_1, \dots, i_k$ on graph $G$;
    **for** $r = 1$ **to** $k$ **do**
        $N_{dy}(x) \leftarrow N_{dy}(x) + d(\hat{y}_i, \hat{y}_{i_r}) \times \delta(i, i_r)$;
        $N_{dx}(x) \leftarrow N_{dx}(x) + d(x_i, x_{i_r}) \times \delta(i, i_r)$;
        $N_{dy\&dx}(x) \leftarrow N_{dy\&dx}(x) + d(\hat{y}_i, \hat{y}_{i_r}) \times d(x_i, x_{i_r}) \times \delta(i, i_r)$;
    **end for**
**end for**
$s(x) \leftarrow \frac{N_{dy\&dx}(x)}{N_{dy}(x)} - \frac{N_{dx}(x) - N_{dy\&dx}(x)}{m - N_{dy}(x)}$;

## 2.4 Application of ML-PI to Glioblastoma Patient Cohort

### 2.4.1 Accuracy on Prediction of Biopsy Samples

Before applying ML-PI, a graph was constructed for each patient/tumor (called target patient hereafter). Vertices of the graph correspond to boxes placed on the T2W ROI of the selected slice for the target patient as well as biopsy samples from other patients. The ML-PI model includes three main parameters that need to be tuned: $\gamma_A$, $\gamma_I$, and $\eta$. The tuning parameter $\eta$ is the width of the radial basis function kernel, $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\eta^2}$. The tuning ranges used were $\gamma_I, \gamma_A \in \{10^{-10}, \dots, 10^4\}$; $\eta \in \{10^{-1}, \dots, 10^2\}$. I compared two tuning strategies: patient-specific tuning and uniform tuning. The former finds the optimal tuning parameters for each patient while the latter

assumes the same optimal tuning parameters across all patients. Specifically, in patient-specific tuning, I trained an ML-PI model for each patient using the augmented biopsy samples from other patients in the loss term. No real or virtual biopsy samples from the target patient were used in training in order to avoid overfitting. Then, the trained model was used to predict the real biopsy samples of the target patient. The optimal tuning parameters were those that minimized the mean absolute prediction error (MAPE) of the target patient. In uniform tuning, I looked for a single set of tuning parameters that minimized the MAPE across all patients. Theoretically, uniform tuning should perform no better than patient-specific tuning. This experiment aimed to find out to what extent patient difference would cause difference in the optimal tuning parameters of ML-PI. Table 1 shows the comparison result using two metrics: MAPE and Pearson correlation between the predicted and pathological cell density measurements. Both metrics considered a 5% error margin for the pathological measurement, i.e., if a predicted value is within $\pm 5\%$ of the pathological measurement, the prediction is considered correct (i.e., with zero prediction error). An MAPE of 0.106 means that if the pathologically measured density of a sample is $\%\ (0 \leq b \leq 100)$, the predicted density by ML-PI deviates from $b\%$ by 10.6% on average. From Table 1 it is clear to see that patient-specific tuning has a significantly better accuracy than uniform tuning in terms of both a smaller MAPE ($p<0.0025$) and a higher Pearson correlation ($p<0.001$).

Table 1: Prediction accuracy of ML-PI: patient-specific and uniform tuning

| | **Patient-specific tuning** | **Uniform tuning** |
|---|---|---|
| **MAPE** | 0.106 ± 0.125 | 0.176 ± 0.177 |
| **Pearson correlation** | 0.838 | 0.588 |

Furthermore, I investigated which of the three tuning parameters have a greater effect on model accuracy when allowed to be patient-specific. To achieve this purpose, I added a third tuning strategy, partially-uniform tuning, in which two of the three tuning parameters were kept the same across all patients while the remaining one was allowed to vary from patient to patient. This results in three models correspond to $\gamma_A$, $\gamma_I$, or $\eta$ as the parameter allowed to be patient-specific, respectively. Table 2 shows the performance of the three models. Compared with the result of uniform tuning in Table 1, it is clear that allowing patient-specific tuning of $\gamma_A$ resulted in a significantly improved MAPE and Pearson correlation ($p = 0.023$ and $0.011$). Patient-specific tuning of $\eta$ does not result in a significantly improved MAPE and Pearson correlation, however the improvement of MAPE approaches the 0.05 significance threshold ($p = 0.087$ and $0.17$). Patient-specific tuning of $\gamma_I$ does not significantly improve the MAPE and Pearson correlation ($p = 0.22$ and $0.35$). Compared with the result of patient-specific tuning of all three parameters in Table 1, patient-specific tuning of $\gamma_A$ alone does not significantly deteriorate the performance in terms of MAPE and Pearson correlation ($p = 0.14$ and $0.39$), while patient-specific tuning of $\eta$ alone shows a greater difference in MAPE and Pearson correlation ($p = 0.057$ and $0.044$) and $\gamma_I$ exhibits a significant deterioration in MAPE and Pearson correlation ($p = 0.012$ and $0.014$). These results show that $\gamma_I$ (and, to some extent, $\eta$) requires less sensitive tuning between patients, suggesting that the Laplacian matrix that incorporates PI similarities successfully accounts for patient differences (thus not necessitating the need for a patient-specific $\gamma_I$).

Table 2: Prediction accuracy of ML-PI with partially-uniform tuning

| | Parameter allowed to be patient-specific | | |
|---|---|---|---|
| | $\gamma_A$ | $\gamma_I$ | $\eta$ |
| **MAPE** | 0.127 ± 0.129 | 0.156 ± 0.154 | 0.140 ± 0.153 |
| **Pearson correlation** | 0.792 | 0.676 | 0.713 |

Next, I compared the performance of ML-PI with PI and ML used alone. The ML model is a supervised learning model that takes the same form of ML-PI except with $\gamma_I = 0$, i.e., a model that does not leverage unlabeled data. Table 3(a) shows the MAPE and Pearson correlation of each model. Compared with ML-PI with patient-specific tuning of all parameters, PI and ML alone had a significantly worse accuracy in terms of both MAPE and Pearson correlation ($p < 0.001$ in all comparisons). Also, I present the patient-wise MAPEs of ML-PI, PI, and ML in Table A1, found in the supplementary information, to allow for comparison on the patient-level. ML-PI was able to predict more accurately than ML and PI in 17 out of 18 patients.

Table 3: Prediction accuracy of ML-PI, PI, and ML on (a) all samples and (b) BAT samples

| **(a)** | **ML-PI** | **PI** | **ML** |
|---|---|---|---|
| **MAPE** | 0.106 ± 0.125 | 0.227 ± 0.215 | 0.199 ± 0.186 |
| **Pearson correlation** | 0.838 | 0.437 | 0.518 |
| **(b)** | **ML-PI** | **PI** | **ML** |
| **MAPE** | 0.132 ± 0.118 | 0.204 ± 0.204 | 0.233 ± 0.209 |
| **Pearson correlation** | 0.820 | 0.416 | 0.208 |

Prediction on the BAT is critically important and challenging. Therefore, I further compared the performance of ML-PI, PI, and ML on samples in the BAT. Out of the 82 total samples, 33 are in this area. Table 3(b) shows the MAPE and Pearson correlation of each model and Figure A3, found in the supplementary information, additionally shows

the predicted vs. pathological cell density for the 33 samples. ML-PI significantly outperforms PI and ML in all the comparisons ($p < 0.05$). Figure 2 additionally shows the predicted vs. pathological cell density for all the samples and the 33 BAT samples in the patient-wise ML-PI, PI, and ML models.

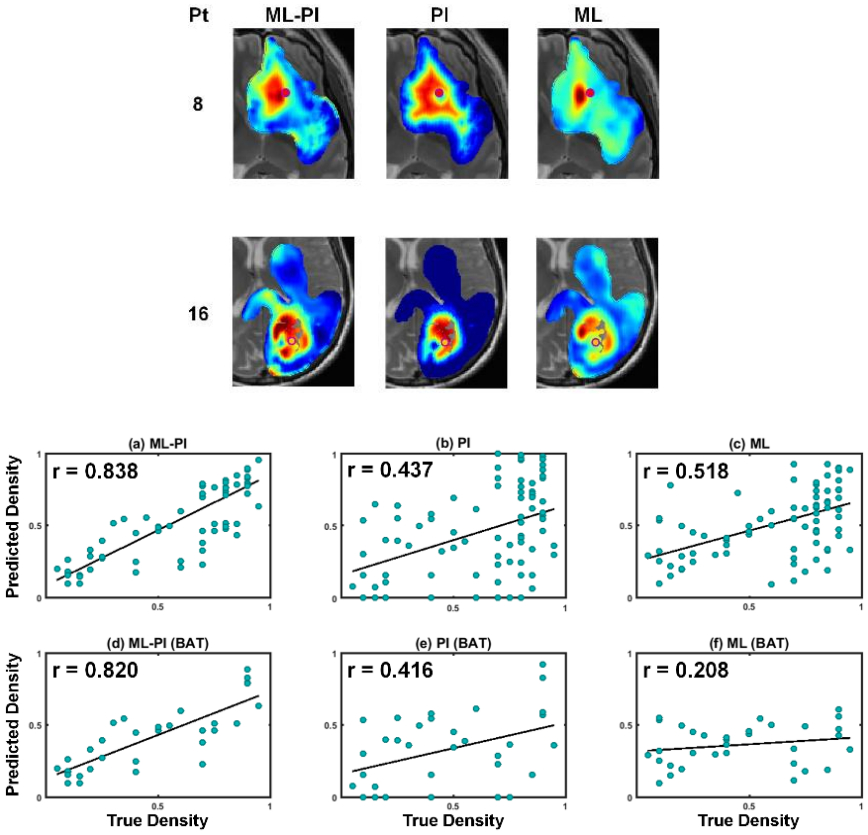



Figure 2: Predicted cell density maps of selected patients and scatter plots of ML-PI, PI, and ML predictions. The left side of the figure shows predicted cell density maps overlaid on T2W image for patients 8 and 16 by three different models. Red to blue colors represent 100%-0% density. A pink circle indicates location of a biopsy sample. For patient 8, the pathological density of the biopsy is 90% and predicted densities by ML-PI, PI, and ML are 79.0%, 59.2%, and 56.4%, respectively. For patient 16, the pathological density of the biopsy is 70% and predicted densities by ML-PI, PI, and ML are 79.4%, 82.9%, and 54.9%, respectively. Below each predicted cell density map are the corresponding patient-wise histograms of the predicted cell densities in the non-enhancing (BAT) regions. The right side of the figure shows predicted density by (a) ML-PI, (b) PI, and (c) ML against pathological density for 82 biopsy samples; predicted density by (d) ML-PI, (e) PI, and (f) ML against pathological density for 33 biopsy samples in non-enhancing (BAT) region. Note that the red and purple boxes indicate the corresponding biopsies shown in the predicted cell density maps for patients 8 and 16 respectively. Additionally, r denotes the Pearson correlation coefficient.

2.4.2 Use of the Trained Model to Generate Whole-Tumor Predicted Density Maps

Ultimately, I would like to generate a predicted cell density map for the T2W ROI in order to guide neurosurgery and radiation therapy. In this experiment, I used the trained

ML-PI model in the previous section to predict tumor cell density on every 8x8 voxel box placed one pixel apart on the T2W ROI. This generated a predicted density map on the T2W ROI. PI can also generate such a map. I compared the maps by ML-PI and PI by generating a patient-wise histogram on predicted density at the BAT. The histograms are shown in Figures A2 and A3, found in Appendix A. It is clear that PI predicted the vast majority of the non-enhancing area to be low density. This is indeed a fundamental assumption of PI. In contrast, ML-PI was able to predict a wider spread of density making it possible to capture high-density regions in the BAT.

Furthermore, I show the predicted cell density maps over the T2W ROIs for two patients in Figure 2. For comparison, maps were predicted by the ML-PI, PI, and ML models for each patient. Pink circles indicate the location where biopsy samples were taken. It can be observed that the map by ML-PI conforms to the global shape of the PI map, and meanwhile predicts more accurately than using PI and ML alone.

### 2.4.3 Contributions from MRI sequences and PI

Using Relief-ML-PI, I can compute a contribution score for each image feature (one feature per MRI sequence) and PI from the ML-PI model specific to each patient. To identify the contributions aggregated over all the patients, I normalize the score of each feature within each patient to be between 0 and 1 by dividing the score by a sum over the scores of all the features. Then, the normalized scores from each patient are added together to produce an aggregated score showing contribution from each feature. Figure 3 shows the contribution from each MRI sequence and PI. It is clear that PI contributes the most.

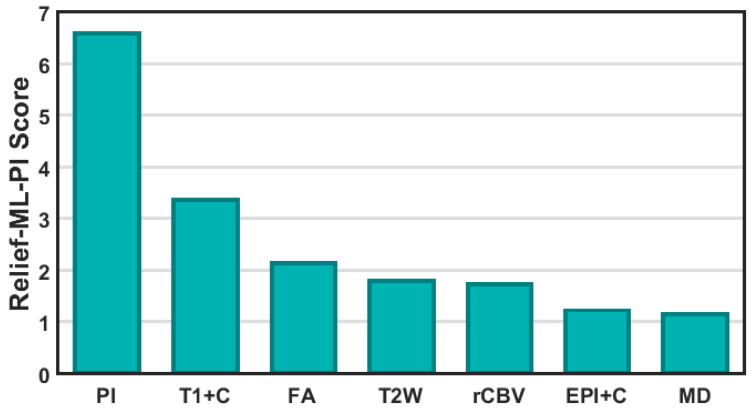



Figure 3: Contributions of PI and MRI sequences to ML-PI cell density prediction.

## 2.5 Conclusion

In this chapter I proposed the ML-PI model that used multiparametric MRI and PI and to regularize tumor cell density prediction under a graph-based SSL framework. ML-PI had capabilities of learning patient-specific relationships between imaging features and cell density, and was found to have a greater prediction accuracy than ML or PI alone when applied to a GBM patient cohort. Additionally, ML-PI showed a more balanced prediction in the T2W ROIs when compared to PI, while the latter underestimated the cell density, indicating that ML-PI was more capable of capturing high density regions in BAT. An algorithm called Relief-ML-PI was also proposed to determine contributions of each individual feature to ML-PI prediction. It was found that PI contributed most significantly to the prediction. This highlighted the importance of incorporating mechanistic models in the form of PI to help improve tumor cell density prediction.

The present study has several limitations. The proposed ML-PI only considers the mechanistic model, PI, as a regularizer in a two-dimensional fashion, whereas PI prediction

is derived as a waveform in three-dimensional (3-D) space. Further work will be performed to incorporate PI as a 3-D regularizer in the model, as this should better utilize the PI prediction and potentially improve the results. Additionally, due to the small number of biopsy samples that can be collected from each patient, an ML-PI active sampling method can be developed to determine optimal locations to sample the tumor before a surgeon collects the stereotactic biopsies.

# CHAPTER 3

# A CLINICAL DECISION SUPPORT SYSTEM USING MULTI-MODALITY IMAGING DATA FOR DISEASE DIAGNOSIS

## 3.1 Background

Imaging has become an indispensable part of modern medicine, and is being extensively used to support diagnosis and other clinical decision making on various diseases such as brain diseases, cardiovascular diseases, and cancer. With the rapid advance of imaging technologies, it is now possible to acquire multiple modalities of imaging data for the same patient. These modalities consist of different but complementary information about the organ of interest, providing an opportunity for better clinical decision support. Taking brain diseases as an example, such as migraine and Alzheimer's disease (AD), a number of imaging modalities can be acquired, which can be broadly classified into structural imaging and functional imaging. Typical structural imaging modalities include computed tomography (CT) and magnetic resonance imaging (MRI): CT shows the gross structure of the brain based on differential absorption of X-rays. MRI produces detailed structural images of the brain using magnetic field and radio waves. Typical functional imaging modalities include functional MRI (fMRI), positron emission tomography (PET), and magnetoencephalography (MEG): fMRI measures blood oxygenation related to neural activity. PET measures physiologic functions in the brain by measuring radiation emitted from tracers injected in the bloodstream. MEG measures magnetic fields produced by the brain's electrical activity using superconducting quantum interference devices.

Recognizing the importance of combining multi-modality imaging data to support disease diagnosis, extensive research has been done, which can be generally categorized into data fusion and data integration. The former interrogates the covariation between different imaging modalities, facilitating knowledge discovery and understanding of the disease biophysiology (Groves et al. 2011, Sui et al. 2011, Calhoun et al. 2006). However, it does not directly support the diagnosis of each individual patient. Data integration aims at utilizing the different but complementary information contained in the multiple imaging modalities in order to assist with disease diagnosis. Methods for data integration share a common idea of building a classifier that links a combined set of features from individual imaging modalities with the diagnostic result. Commonly used classification models include linear discriminant analysis (LDA) (Huang et al. 2011, Hu et al. 2015), quadratic discriminant analysis (QDA) (Schwedt et al. 2015, Chong et al. 2017, Zhang et al. 2016), support vector machines (SVM) (Fan et al. 2008, Yang et al. 2010, Zhang et al. 2011), and multitask learning (Yu et al. 2014, Yuan et al. 2012). Integrating multi-modality imaging data has been shown to produce better classification accuracy than using a single modality alone in a number of diseases such as AD (Huang et al. 2011, Fan et al. 2008, Zhang et al. 2011, Yu et al. 2014, Yuan et al. 2012), schizophrenia (Yang et al. 2010), migraine (Chong et al. 2017, Schwedt et al. 2015), and glioblastoma (Hu et al. 2015, Hu et al. 2016).

Despite the abundance of existing research, the research has not yet been transformed into a clinical decision support system due to the lack of three important traits: *flexibility*, *sufficient accuracy*, and *interpretability*. *Flexibility* means that the system can incorporate image features defined at various aggregation levels such as voxels and regions of interest (ROIs). Both voxel-level and ROI-level features are commonly used in imaging-

based studies and have their respective strengths: The former preserves the raw information in an image, which avoids information loss. The latter combines prior knowledge (e.g., the anatomical structure of an organ) to guide feature definition. Furthermore, a system with *flexibility* should be able to take image features of various types such as element (voxel or ROI)-wise features and connectivity-based features. Examples of element-wise features include cortical thickness, area, and volume using MRI and regional metabolism using PET. Examples of connectivity-based features include functional connectivity z-maps using fMRI and white matter tractography using diffusion tensor imaging (DTI). Lastly, most multi-modality imaging based studies require co-registration to ensure the images are aligned into the same coordinate system (Maintz and Virgever 1998, Hajnal and Hill 2001), which is time consuming and error-prone. A system with *flexibility* should provide an option for opting out this procedure.

*Sufficient accuracy* means a superior performance of the classification model which can be used for individual patient diagnosis instead of group-based analysis. Given the high-dimensionality of the joint feature set produced by multi-modality images, searching for the subset of features with the near-global optimal classification accuracy is very challenging. An exhaustive search is practically impossible. Greedy search based methods such as sequential forward selection and sequential backward selection suffer from a variety of problems such as stagnation in local optima and a high computational cost. Lately, evolutionary computation (EC) techniques such as genetic algorithms (GA) (Fraser and Burnell 1970), genetic programming (GP) (Koza 1990), differential evolution (DE) (Storn and Price 1997), and neuroevolution (Floreano et al. 2008) have attracted great attention with some initial success in feature selection and classification for medical

applications. A new emerging field in EC is swarm intelligence (Bonyadi and Michalewicz 2017) which models the collective behavior of social swarms in nature, such as ant colonies, honeybees, and bird flocks. Although individuals in a swarm are relatively unsophisticated with limited capabilities on their own, they interact together with certain behavioral patterns to cooperatively achieve tasks necessary for their survival. This "intelligent" behavior of the swarm has inspired new algorithmic developments in solving large complex optimization problems with a wide range of application domains such as machine learning (Das et al. 2009), bioinformatics (Das et al. 2008), dynamical systems and operations research (Parsopoulos 2010). Particle swarm optimization (PSO) is a computational algorithm based on swarm intelligence that mimics the behavior of flying birds and their means of information exchange to solve optimization problems. Each potential solution is seen as a particle with a certain velocity, and flies through the problem space. Each particle adjusts its flight according to its own flying experience and its companions' flying experiences. The particle swarms find optimal regions over complex search spaces through the interaction of individuals in a population of particles. PSO has been successfully applied to a number of difficult combinatorial optimization problems (Jarboui et al. 2008, Chu et al. 2012). PSO has also been shown to be computationally less expensive, converge more quickly, and find better solutions than classic EC algorithms such as GA (Wang et al. 2007, Jarboui et al. 2007).

*Interpretability* is another important trait that a clinical decision support system should possess. In general, mathematical models can be described as black-box, white-box, or grey-box (Khan and Khan 2012). Black-box models do not convey information about their inner-workings, and only the input and output are known. White-box models convey

explicit information about their internal structure, allowing the user to infer the different components and their connections. Grey-box models display partial theoretical information and use the data that is available to complete the model. In this research, white-box approaches in feature processing and model building are employed to achieve interpretability as it would allow for identification of an analytic pathway that traces back from the classification accuracy to the contributing features and their respective contributing weights. This has at least two benefits: First, it facilitates identification of biomarkers for the disease. Biomarker identification is of vital importance in medical research not only for disease diagnosis but also for understanding the biological basis and developing effective treatments. Second, practitioners tend to be reluctant to adopt black-box approaches regardless of the performance. White-box approaches allow for ready clinical adaptation and dissemination.

In this research, I develop a multi-modality imaging based diagnostic decision support system (MMI-DDS) aiming to possess the aforementioned three traits. MMI-DDS includes three key steps: First, a modality-wise principal component analysis (PCA) is applied to each imaging modality independently. Imaging features are typically high-dimensional. Some features are naturally highly correlated due to their spatial proximity or functional similarity. These pose challenges to downstream classification model development. PCA is a well-known statistical method for dimension reduction and de-correlation. PCA is also a white-box approach because it applies a linear transformation to the imaging features, which allows for a later inverse-transformation to identify the contributing features to the classification accuracy (i.e., the biomarkers). In MMI-DDS, a modality-wise PCA is employed in order to account for the fact that different imaging

modalities may measure the organ of interest from different perspectives. This also provides an option for opting out tedious and error-prone co-registration for the multi-modality images. Second, a novel constrained PSO (cPSO) based classifier is built on the joint set of principal components (PCs) across the multi-modalities. cPSO is an optimizer that searches through the joint PC set to find a small subset of PCs with near-global optimal classification accuracy. In this sense, cPSO combines feature (i.e., PCs) selection and classification in a single framework. The ability of feature selection is important for medical applications since medical data tend to contain many features. Simply training a classifier to all the available features would likely cause overfitting since many of the features are likely to be noise. In theory, the cPSO optimizer can be used for all classification models. In this chapter, I choose white-box models such as LDA, QDA, and linear SVM (LSVM) to enable inverse-transformation and biomarker identification in the next step. Third, a clinical utility engine is developed to derive the analytic pathway that traces back from the classification accuracy to the contributing features (i.e., biomarkers) and their respective contributing weights. This allows for interpretation of the diagnostic result and knowledge discovery about the disease.

The rest of the chapter is structured as follows: Section 3.2 provides a literature review. Section 3.3 presents development of the MMI-DDS. Section 3.4 presents an application of MMI-DDS in using multiparametric MRI to predict intra-tumor heterogeneity. Section 3.5 presents an application of MMI-DDS for migraine diagnosis using multi-modality structural and functional imaging data. Section 3.6 is the conclusion.

## 3.2 Literature Review

As mentioned in the Section 3.1, research on combining multi-modality imaging data falls into two categories: data fusion and data integration. This chapter belongs to the latter category, but I will review the existing work in both categories in this section due to their relevance.

For data fusion, multivariate statistical methods such as canonical correlation analysis (CCA), partial least squares (PLS), and independent component analysis (ICA) provide viable approaches. CCA finds linear combinations of two sets of variables, called canonical variables, with the maximum correlation between each other. The original CCA can only model two datasets. It was later extended to a multiset-CCA (M-CCA) that finds canonical variables from multiple datasets to achieve the maximum overall correlation (Kettenring 1971). M-CCA was used to perform data fusion of concurrently acquired fMRI and EEG in an auditory task to find covarying amplitude modulations in both modalities and the corresponding spatial activations (Correa et al. 2010). It was also used to fuse fMRI, EEG, and MRI to make group inference for schizophrenia patients compared with healthy controls (Correa et al. 2009).

PLS is a statistical model that finds the multidimensional direction in the space of the independent variables that explains the maximum multidimensional variance direction in the space of the dependent variables. Multiway PLS, as an extension to PLS, was developed for fusion of EEG and fMRI by decomposing EEG and fMRI each as a sum of "atoms" (Martinez-Montes et al. 2004). Each EEG atom was the outer product of spatial, spectral, and temporal signatures and each fMRI atom the product of spatial and temporal signatures. The decomposition was constrained to maximize the covariance between

corresponding temporal signatures of the EEG and fMRI. This fusion aimed at identifying the coherent systems of neural oscillators that contribute to the spontaneous EEG.

ICA is a generative model that assumes the observed multivariate data to be weighted sums of unobserved independent components. ICA is a popular approach in image analysis. Earlier work focused on single imaging modalities such as fMRI and EEG with the purpose of separating the imaging data into meaningful constituent components correlated with subjects' experimental task performance. Recently, ICA has been extended in a number of ways for multi-modality data fusion. Joint ICA (jICA) assumes that the data from multiple imaging modalities share a common demixing matrix (Calhoun and Adali 2009). Several studies demonstrated the use of jICA in fusion of fMRIs from multiple tasks, MRI and fMRI, fMRI and EEG, and MRI and DTI for identifying group difference between patients with schizophrenia and controls (Calhoun et al. 2006, Calhoun and Adali 2009, Xu et al. 2009]. Parallel ICA (paraICA) (Sui et al. 2011, Calhoun and Adali 2009, Liu et al. 2009) was developed to relax the strong "common demixing matrix" assumption posed by jICA and provided a more flexible approach by creating the mixing matrices for different modalities separately with the goal of maximizing the independence of components within each modality while maximizing the correlation between the mixing matrices. paraICA was used to fuse fMRI and SNP (a genetic modality) in studying schizophrenia (Liu et al. 2009) and to fuse fMRI and DTI in comparing schizophrenia with bipolar disorder (Sui et al. 2011). Tensor ICA (Beckmann and Smith 2005) was developed to fuse three-way (spatial, temporal, and cross-subject) fMRI data by decomposing the data into a set of independent spatial maps together with associated time courses and estimated subject modes. It was applied to fMRI data collected under a visual, cognitive, and motor

paradigm and was able to extract plausible activation maps, time courses, and session/subject modes as well as provide a rich description of additional processes of interest such as image artifacts and secondary activation patterns. Link ICA adopted a Bayesian framework for simultaneously modeling and discovering common features across multiple modalities (Groves et al. 2011). It enjoyed the flexibility of fusing imaging modalities with completely different units, signal- and contrast-to-noise ratios, voxel counts, spatial smoothness and intensity distributions by using a Bayesian formulation to automatically weigh the modalities appropriately.

While being a popular research area, multi-modality imaging data fusion does not directly support diagnosis of each individual patient, but instead provides an exploratory tool for knowledge discovery and group inference. The former is the objective of multi-modality imaging data integration. Research on data integration shares a common idea of building a classifier from a training dataset, which links a combined set of features from individual imaging modalities with a diagnostic result. This classifier can then be used to produce a probability of having the target disease for each new patient, thus providing decision support for clinical diagnosis. In theory, such a classifier can be built using any statistical classification method. Typical methods that have been used for integrating multi-modality imaging data include LDA (Huang et al. 2011, Hu et al. 2015), QDA (Schwedt et al. 2015, Chong et al. 2016), SVM (Fan et al. 2008, Yang et al. 2010, Zhang et al. 2011), and multitask learning (Yu et al. 2014, Yuan et al. 2012). Integrating multi-modality imaging data has been shown to produce better classification accuracy than using a single modality alone in a number of brain diseases such as AD (Huang et al. 2011, Fan et al. 2008, Zhang et al. 2011, Yu et al. 2014, Yuan et al. 2012), schizophrenia (Yang et al. 2010),

migraine (Schwedt et al. 2015, Chong et al. 2017), and glioblastoma (Hu et al. 2015, Hu et al. 2016). Despite the abundance of existing literature, the research is still limited in clinical usability due to lack of flexibility (e.g., only applicable to certain imaging modalities or requiring co-registration), insufficient accuracy (e.g., using off-the-shelf software to build a classification model without exploiting advanced optimizers to improve the performance), and insufficient interpretability (e.g., black-box methods prohibiting rigorous identification of contributing features or biomarkers).

## 3.3 Development of MMI-DDS

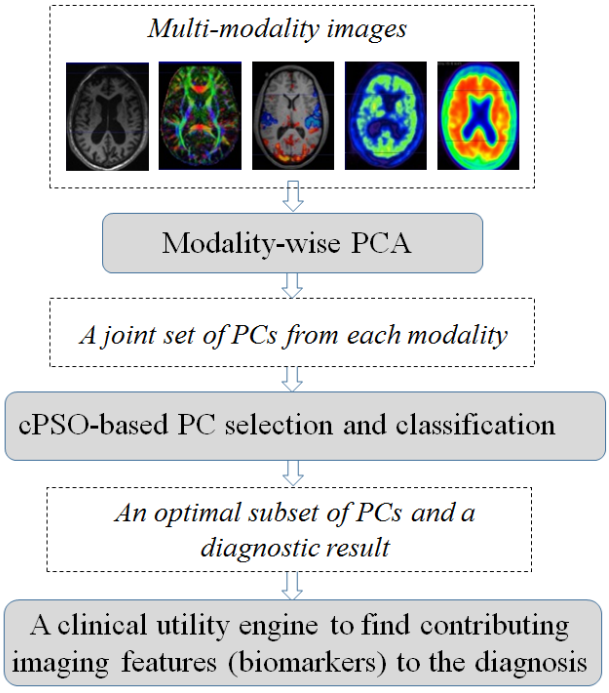



Figure 4: Layout of MMI-DDS

As shown in Figure 4, MMI-DDS includes the following main components: (1) a modality-wise PCA, (2) a cPSO-based classifier for diagnosis, and (3) a clinical utility engine for biomarker identification.

### 3.3.1 Modality-wise PCA

PCA is a statistical method that transforms the imaging features that are potentially high-dimensional and correlated into a small number of uncorrelated PCs. Each PC is a linear combination of the imaging features. The transformation is performed in such a way that the first PC has the largest possible variance and each succeeding PC has the highest variance possible under the constraint that it is uncorrelated with all the preceding PCs. I propose to perform PCA on each imaging modality separately. This is to account for the fact that different imaging modalities measure the organ of interest from different perspectives and therefore combining their features in a single PCA is inappropriate. This also provides the flexibility for opting out co-registration of the multi-modality images. Specifically, suppose there are $M$ imaging modalities. Let $\mathbf{X}_m = [X_{1,m}, \dots, X_{n_m,m}]^T$ be the set of features corresponding to the $m$-th modality, $m = 1, \dots, M$. $n_m$ is the number features for the $m$-th modality. Let $\mathbf{Z}_m = [Z_{1,m}, \dots, Z_{p_m,m}]^T$ be the set of PCs. Each PC is a linear combination of the features, i.e., $Z_{i,m} = \mathbf{w}_{i,m}^T \mathbf{X}_m$. $\boldsymbol{w}_{i,m}$ consists of the combination coefficients and is called the loading vector. To obtain the loading vectors for all the PCs, a dataset on the features $\mathbf{X}_m$ needs to be collected, which consists of measurements on $\mathbf{X}_m$ from $N$ samples (i.e., patients). Using the dataset, a sample correlation matrix of $\mathbf{X}_m$, $\mathbf{S}_m$, can be computed and an eigen-decomposition is further performed on $\mathbf{S}_m$. The eigenvalues will be ordered from the largest to the smallest, $\lambda_{1,m}, \dots, \lambda_{p_m,m}$, and the corresponding eigenvectors are the loading vectors for the first through the last PC. Note that not all the PCs need to be kept for subsequent analysis, since the PCs corresponding to small eigenvalues are likely to capture noise in the data but not useful information. To determine

the number of PCs to keep, a typical approach is to keep track of the cumulative percentage of variance explained by adding more PCs until a pre-specified threshold is reached. Setting the threshold to be a number between 80%-90% has been found to be adequate for most applications (Hu et al. 2015, Schwedt et al. 2015, Chong et al. 2017).

### 3.3.2 cPSO-based Feature Selection and Classification

PSO was originally developed as a population-based stochastic optimization technique, and then extended for feature selection in classification. In this section, I first briefly introduce how generic PSO works for solving an optimization problem and for feature selection. Then, I propose a modified PSO algorithm that can honor a pre-specified maximum number of features to better avoid overfitting, called cPSO.

Consider an optimization problem with decision variables $x_1, \dots, x_D$ and an objective function $f(x_1, \dots, x_D)$ to optimize. PSO is initialized with a population of random solutions called particles. Let $\mathbf{x}_i = (x_{i1}, \dots, x_{iD})$ represent the $i$-th particle, $i = 1, \dots, I$. Each particle adjusts its position according to its own experience and the positions of other particles. Specifically, at the $t$-th iteration, let $\mathbf{p}_i^t$ be the best previous position of the $i$-th particle (i.e., the position giving the best value for the objective function) and $\mathbf{p}_g^t$ be the best position among all the particles. Then, the position adjustment, called velocity, of the $i$-th particle along the $d$-th dimension is given by:

$$v_{id}^t = \omega^t v_{id}^{t-1} + c_1 r_1 (p_{id}^t - x_{id}^t) + c_2 r_2 \left(p_{gd}^t - x_{id}^t\right), \tag{3.1}$$

$d = 1, \dots, D$. Here, $\omega^t$, $c_1$, and $c_2$ are called the inertia weight, cognitive learning factor, and social learning factor, respectively. A proper choice for $\omega^t$ provides a balance between

global and local exploration, and results in fewer iterations on average to find a sufficiently optimal solution. $c_1$ and $c_2$ represent the weighting of the stochastic acceleration terms that pull each particle toward $\mathbf{p}_i^t$ and $\mathbf{p}_g^t$ (Wang et al. 2007). $\omega^t$, $c_1$, and $c_2$ can be treated as tuning parameters of the PSO algorithm. Alternatively, they can be set by users. A number of appropriate values for the three parameters have been suggested (Poli et al. 2007). $r_1$ and $r_2$ are sampled from a uniform distribution $U[0,1]$. Furthermore, according to the velocity in (1), the $i$-th particle can move to a new position, i.e.,

$$x_{id}^{t+1} = x_{id}^t + v_{id}^t. \tag{3.2}$$

Kennedy and Eberhardt proposed modifications on the afore-described generic PSO, so that the resulting algorithm can be used for feature selection in classification (Kennedy and Eberhart 1997). Suppose there are $D$ features, $Z_1, \dots, Z_D$. Each feature $Z_d$ is associated with a binary decision variable $x_d$. $x_d = 1$ if $Z_d$ is selected and $x_d = 0$ otherwise. The objective function $f(x_1, \dots, x_D)$ is a cross-validated classification error that is computed using the selected features on a training dataset. Because of the binary nature of the decision variables, (3.2) is changed to (3.3) while (3.1) remains the same.

$$x_{id}^{t+1} = \begin{cases} 1, & if \ \ S(v_{id}^t) > r \\ 0, & otherwise \end{cases}, \tag{3.3}$$

where $S(v_{id}^t)$ is a sigmoid function used to map $v_{id}^t$ to [0,1], i.e., $S(v_{id}^t) = \frac{1}{1+e^{-v_{id}^t}}$. $r$ is sampled from $U[0,1]$.

In this chapter, I propose a cPSO algorithm that can honor a pre-specified maximum number of features to avoid overfitting. Specifically, I modify (3.3) as follows: Let $K$ denote the maximum number of features allowed in the classification model. For each particle, I order its velocities along all the dimensions from the largest to the smallest.

Without loss of generality, I denote the ordered velocities of the $i$-th particle by $v_{i1}^t, \dots, v_{iD}^t$. Keep the first $K$ largest velocities, $v_{i1}^t, \dots, v_{iK}^t$. A simple modification on (3.3) could be to make $x_{id}^{t+1} = 1$ if $d \leq K$ and $x_{id}^{t+1} = 0$ otherwise. Although this approach guarantees $K$ features to be selected, the selected features may have poor quality. Here, I consider a feature to have poor quality if it has a negative velocity, $v_{id}^t < 0$, which leads to the sigmoid function $S(v_{id}^t) < 0.5$. Therefore, (3.3) is modified into (3.4) in cPSO:

$$x_{id}^{t+1} = \begin{cases} 1, & if \;\; d \leq K \; and \; S(v_{id}^t) > 0.5 \\ 0, & otherwise \end{cases}. \tag{3.4}$$

Using (3.4), only the $K$ largest features that have good quality, i.e., have a higher probability of being selected than not being selected, will be kept. Therefore, the number of selected features can be less than or equal to $K$.

Next, I present the detailed steps of the cPSO algorithm. The input to cPSO includes a training dataset on the joint set of PCs by pooling together the PCs from each imaging modality, denoted by $Z_1, \dots, Z_D$, and a diagnostic result $Y$. The input also includes several user-specified parameters: the maximum number of PCs, $K$; the number of particles, $I$; the number of iterations, $T$; the maximum velocity used to limit further exploration after convergence to an optimal value, $V_{max}$. Set $\omega^t = 0.9 - t \cdot 0.5/T$, $c_1 = 2$, and $c_2 = 2$, which are recommended values by the literature (Poli et al. 2007). In addition, a classification model needs to be specified. In theory, cPSO can work with any classification model. In this chapter, I focus on white-box models such as LDA, QDA, and LSVM. This is to facilitate identification of the contribution features (i.e., biomarkers) and their respective contributing weights to the classification accuracy in a mathematically and computationally tractable way.

***The proposed cPSO algorithm:***

***Step 1 (initialization):*** Set the initial position of the $i$-th particle, $\mathbf{x}_i^0$, by randomly choosing $K$ elements in $\mathbf{x}_i^0$ to be one while making other elements to be zero. Use the PCs corresponding to the non-zero elements in $\mathbf{x}_i^0$ to compute a cross validated (CV) classification error on the training dataset, $f(\mathbf{x}_i^0)$. Set the initial velocity, $\mathbf{v}_i^0$, by sampling each element in $\mathbf{v}_i^0$ from $U[-V_{max}, V_{max}]$. Use (4) to update the initial position of each particle and get $\mathbf{x}_i^1$. Iterate Steps 2-3 with $t = 1,2, \dots T$.

***Step 2 (velocity updating):*** Examine all previous positions of the $i$-th particle, $f(\mathbf{x}_i^0), \dots, f(\mathbf{x}_i^{t-1})$, and find the position giving the smallest CV classification error, $\mathbf{p}_i^t$. Examine the current positions of all the particles, $f(\mathbf{x}_1^t), \dots, f(\mathbf{x}_I^t)$ and find the position giving the smallest CV classification error, $\mathbf{p}_g^t$. Sample $r_1$ and $r_2$ from $U[0,1]$. Use (1) to compute the velocity $\mathbf{v}_i^t$. If $v_{id}^t > V_{max}$, set $v_{id}^t = V_{max}$; if $v_{id}^t < -V_{max}$, set $v_{id}^t = -V_{max}$.

***Step 3 (position updating):*** Order the elements in $\mathbf{v}_i^t$ from the largest to the smallest. Use (4) to compute the new position $\mathbf{x}_i^{t+1}$. If the maximum number of iterations has been reached, i.e., $t + 1 = T$, examine the current positions of all the particles, $f(\mathbf{x}_1^{t+1}), \dots, f(\mathbf{x}_I^{t+1})$, and output the position giving the smallest CV classification error as the optimal solution, together with the corresponding CV error and the PCs that are selected. Otherwise, go back to Step 2.

Finally, I discuss how to select the maximum number of PCs, $K$. A general trend is that the CV classification error will decrease as $K$ increases. However, this does not mean that a larger $K$ is always preferred, because the decrease in the CV error after $K$ is beyond a certain value is so minimal that it is neither statistically significant nor practically useful.

Allowing a larger $K$ than needed will produce an over-complicated model that likely has problems with over-fitting. Therefore, a recommended approach for choosing the optimal $K$, i.e., $K^*$, is to plot the CV errors against different values of $K$ with $K$ ranging from the smallest to the largest, and look for the "elbow" point as the $K^*$. This is a similar idea to the scree plot used to find the optimal number of PCs in PCA. Alternatively, a more rigorous approach that uses hypothesis testing may be adopted (e.g., a two-sample t test) to compare the CV errors corresponding to $K$ and $K+1$, $K = 1,2,$ .... The $K^*$ could be one whose CV error is significantly smaller than that of $K^* - 1$ but not than $K^* + 1$. Other methods for choosing $K^*$ might also be adopted, such as penalizing the error with $K$ (similar to the methods used with AIC and BIC). I acknowledge that this is an open area that no single approach dominates. In practice, these alternative approaches could be tried and the results may be cross-referenced with each other.

### 3.3.3 Clinical Utility Engine for Clinical Interpretation and Biomarker Identification

The goal of the clinical utility engine is to identify the contributing original features and their respective contributing weights to the model with best classification accuracy found by cPSO. These can be analytically derived for white-box classification models such as LDA, QDA, and LSVM. I first define some common notations: Let $\mathbf{z}$ be the set of PCs selected by cPSO. $\mathbf{z} = [\mathbf{z}_1^T, \dots, \mathbf{z}_M^T]^T$, where $\mathbf{z}_m$ represents the selected PCs from the $m$-th modality, $m = 1, \dots, M$.

$$\mathbf{z}_m = \mathbf{W}_m^T \mathbf{X}_m \,, \tag{3.5}$$

where $\mathbf{W}_m$ is the loading matrix obtained from the modality-wise PCA discussed in Section 3.1. Let $\mathbf{w}_m^{j\,T}$ be the $j$-th row of $\mathbf{W}_m$. Then, (3.5) can be written as

$$\mathbf{z}_m = \sum_{j=1}^{n_m} \mathbf{w}_m^j X_{j,m}. \quad (3.6)$$

Next, I will present the development of three inverse-operators for LDA, QDA, and LSVM in achieving the goal of the engine.

***LDA inverse operator***

The LDA model takes the following form:

$$log\frac{P(Y=1|\mathbf{z})}{P(Y=0|\mathbf{z})} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}\mathbf{z} - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + log\frac{\pi}{1-\pi}, \quad (3.7)$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$ are the means of $\mathbf{z}$ for the two classes. LDA assumes that the two classes have the same covariance matrix of $\mathbf{z}$, which is represented by $\boldsymbol{\Sigma}$. $\pi = P(Y=1)$. The classification rule of LDA is that if $log\frac{P(Y=1|\mathbf{z})}{P(Y=0|\mathbf{z})} > 0$, assign the sample to class 1, and to class 0 otherwise.

$\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}$, and $\pi$ can be estimated from training data by maximum likelihood estimation (MLE). Then, (3.7) can be simplified as:

$$log\frac{P(Y=1|\mathbf{z})}{P(Y=0|\mathbf{z})} = \mathbf{v}^{\boldsymbol{T}}\mathbf{z} + v_0, \quad (3.8)$$

where $\mathbf{v} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ and $v_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + log\frac{\pi}{1-\pi}$. Letting $\mathbf{v} = [\mathbf{v}_1^T, \dots, \mathbf{v}_M^T]^T$, where $\mathbf{v}_m$ are the coefficients corresponding to $\mathbf{z}_m$, and substituting (3.6) into (3.8), the following is obtained

$$log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})} = \sum_{m=1}^{M}\sum_{j=1}^{n_m}\mathbf{v}_m^T\mathbf{w}_m^j X_{j,m} + v_0. \quad (3.9)$$

It is clear from (3.9) that the magnitude of $\mathbf{v}_m^T \mathbf{w}_m^j$ indicates the contribution of each imaging feature $X_{j,m}$ to the classification accuracy. The sign of $\mathbf{v}_m^T \mathbf{w}_m^j$ indicates the direction of the contribution.

***QDA inverse operator***

The QDA model takes on the following form:

$$log\frac{P(Y=1|\mathbf{z})}{P(Y=0|\mathbf{z})} = -\frac{1}{2}\mathbf{z}^T(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_0^{-1})\mathbf{z} + (\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1})\mathbf{z} - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + log\frac{\pi}{1-\pi} + log\sqrt{|\boldsymbol{\Sigma}_0|/|\boldsymbol{\Sigma}_1|}, \quad (3.10)$$

QDA assumes that the two classes have the different covariance matrices of $\mathbf{z}$, which are represented by $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_0$. Then, (3.10) can be simplified as:

$$log\frac{P(Y=1|\mathbf{z})}{P(Y=0|\mathbf{z})} = \mathbf{z}^T\boldsymbol{\Phi}\mathbf{z} + \mathbf{q}^T\mathbf{z} + q_0, \quad (3.11)$$

where $\boldsymbol{\Phi} = -\frac{1}{2}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_0^{-1})$, $\mathbf{q} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$, and $q_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + log\frac{\pi}{1-\pi} + log\sqrt{|\boldsymbol{\Sigma}_0|/|\boldsymbol{\Sigma}_1|}$. $\boldsymbol{\Phi}$ is a block diagonal matrix under the assumption that the modalities are independent, i.e., $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Phi}_M \end{bmatrix}$, where $\boldsymbol{\Phi}_m$ is the matrix corresponding to the $m$-th modality, $m = 1, \dots, M$. Letting $\mathbf{q} = [\mathbf{q}_1^T, \dots, \mathbf{q}_M^T]^T$, where $\mathbf{q}_m$ are the coefficients corresponding to $\mathbf{z}_m$, and substituting (3.6) into (3.11), I get

$$log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})} = \sum_{m=1}^{M}\sum_{j=1}^{n_m}\mathbf{w}_m^{j\,T}\mathbf{\Phi}_m\mathbf{w}_m^j X_{j,m}^2 + \sum_{m=1}^{M}\sum_{j=1}^{n_m}\mathbf{q}_m^T\mathbf{w}_m^j X_{j,m} +$$

$$\sum_{m=1}^{M}\sum_{j=1}^{n_m}\sum_{\substack{k=1\\k\neq j}}^{n_m}\mathbf{w}_m^{j\,T}\mathbf{\Phi}_m\mathbf{w}_m^k X_{j,m}X_{k,m} + q_0. \qquad (3.12)$$

It is difficult to assess the contribution of each imaging feature $X_{j,m}$ to the classification accuracy based on (3.12), because of the existence of the cross terms $X_{j,m}X_{k,m}$, $k = 1, \dots, n_m, k \neq j$. To tackle this difficulty, I propose to take the expectation of $log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})}$ with respect to the $X_{k,m}$'s, or equivalently the conditional expectation of $log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})}$ with respect to $\mathbf{X}_m$ given $X_{j,m}$. This would average out the contribution from each $X_{k,m}$ and leave only the $X_{j,m}$ to be linked with the classification accuracy. Specifically,

$$E_{\mathbf{X}_m|X_{j,m}}\left[log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})}\right] =$$

$$\mathbf{w}_m^{j\,T}\mathbf{\Phi}_m\mathbf{w}_m^j X_{j,m}^2 + \mathbf{q}_m^T\mathbf{w}_m^j X_{j,m} + \sum_{\substack{k=1\\k\neq j}}^{n_m}\mathbf{w}_m^{k\,T}\mathbf{\Phi}_m\mathbf{w}_m^k \cdot E_{\mathbf{X}_m|X_{j,m}}\left[X_{k,m}^2\right] +$$

$$\sum_{\substack{k=1\\k\neq j}}^{n_m}\mathbf{q}_m^T\mathbf{w}_m^k \cdot E_{\mathbf{X}_m|X_{j,m}}[X_{k,m}] + 2\cdot\sum_{\substack{k=1\\k\neq j}}^{n_m}\mathbf{w}_m^{j\,T}\mathbf{\Phi}_m\mathbf{w}_m^k \cdot X_{j,m}\cdot E_{\mathbf{X}_m|X_{j,m}}\left[X_{k,m}\right] +$$

$$\sum_{\substack{k=1\\k\neq j}}^{n_m}\sum_{\substack{l=1\\l\neq j\\l\neq k}}^{n_m}\mathbf{w}_m^{k\,T}\mathbf{\Phi}_m\mathbf{w}_m^l \cdot E_{\mathbf{X}_m|X_{j,m}}[X_{k,m}X_{l,m}] + q_{0,m} + f(\mathbf{X}_{-m}), \qquad (3.13)$$

where $q_{0,m}$ denotes the portion of $q_0$ that is associated with the $m$-th modality. Since my purpose here is to assess the contribution of $X_{j,m}$, the imaging features from other modalities than the $m$-th modality are not relevant. Therefore, the terms involving these features are put into $f(\mathbf{X}_{-m})$. Furthermore, assume that the imaging features in each modality follows a multivariate normal distribution, i.e., $\mathbf{X}_m \sim N(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, where $\boldsymbol{\mu}_m =$

$\left(\mu_{1,m} \dots \mu_{n_m,m}\right)^T$ and $\boldsymbol{\Sigma}_m = \begin{pmatrix} \sigma_{1,1,m} & \cdots & \sigma_{1,n_m,m} \\ \vdots & \ddots & \vdots \\ \sigma_{n_m,1,m} & \cdots & \sigma_{n_m,n_m,m} \end{pmatrix}$. $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ can be estimated from training data. Under this distribution, the expectations in (3.13) can be derived as:

$$E_{\mathbf{X}_m|X_{j,m}}\left[X_{k,m}\right] = \mu_{k,m} + \frac{\sigma_{k,j,m}\left(X_{j,m}-\mu_{j,m}\right)}{\sigma_{j,j,m}}, \tag{3.14a}$$

$$E_{\mathbf{X}_m|X_{j,m}}\left[X_{k,m}^2\right] = \sigma_{k,k,m} - \frac{\sigma_{k,j,m}^2}{\sigma_{j,j,m}} + \left(\mu_{k,m} + \frac{\sigma_{k,j,m}\left(X_{j,m}-\mu_{j,m}\right)}{\sigma_{j,j,m}}\right)^2, \tag{3.14b}$$

$$E_{\mathbf{X}_m|X_{j,m}}[X_{k,m}X_{l,m}] = \sigma_{k,l,m} - \frac{\sigma_{k,j,m}\sigma_{l,j,m}}{\sigma_{j,j,m}} + \left(\mu_{k,m} + \frac{\sigma_{k,j,m}\left(X_{j,m}-\mu_{j,m}\right)}{\sigma_{j,j,m}}\right)\left(\mu_{l,m} + \frac{\sigma_{l,j,m}\left(X_{j,m}-\mu_{j,m}\right)}{\sigma_{j,j,m}}\right). \tag{3.14c}$$

After substituting (3.14a-c) into (3.13), (3.13) can be simplified to the general form of

$$E_{\mathbf{X}_m|X_{j,m}}\left[log\frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})}\right] = Q_{j,m}\cdot X_{j,m}^2 + L_{j,m}\cdot X_{j,m} + c_{j,m},$$

where $Q_{j,m}$ and $L_{j,m}$ given by:

$$Q_{j,m} = \mathbf{w}_m^{j\,T}\boldsymbol{\Phi}_m\mathbf{w}_m^j + \sum_{\substack{k=1\\k\neq j}}^{n_m}\left(\frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\right)^2\mathbf{w}_m^{k\,T}\boldsymbol{\Phi}_m\mathbf{w}_m^k + 2\cdot\sum_{\substack{k=1\\k\neq j}}^{n_m}\left(\frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\right)\mathbf{w}_m^{j\,T}\boldsymbol{\Phi}_m\mathbf{w}_m^k + \sum_{\substack{k=1\\k\neq j}}^{n_m}\sum_{\substack{l=1\\l\neq j\\l\neq k}}^{n_m}\left(\frac{\sigma_{k,j,m}\sigma_{l,j,m}}{\sigma_{j,j,m}^2}\right)\mathbf{w}_m^{k\,T}\boldsymbol{\Phi}_m\mathbf{w}_m^l, \tag{3.15a}$$

$$L_{j,m} = \mathbf{q}_m^T\mathbf{w}_m^j + \sum_{\substack{k=1\\k\neq j}}^{n_m}\left(\frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\right)\mathbf{q}_m^T\mathbf{w}_m^k + 2\cdot\sum_{\substack{k=1\\k\neq j}}^{n_m}\left(\frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\left(\mu_{k,m} - \frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\mu_{j,m}\right)\right)\mathbf{w}_m^{k\,T}\boldsymbol{\Phi}_m\mathbf{w}_m^k + 2\cdot\sum_{\substack{k=1\\k\neq j}}^{n_m}\left(\mu_{k,m} - \frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\mu_{j,m}\right)\mathbf{w}_m^{j\,T}\boldsymbol{\Phi}_m\mathbf{w}_m^k + \sum_{\substack{k=1\\k\neq j}}^{n_m}\sum_{\substack{l=1\\l\neq j\\l\neq k}}^{n_m}\left(\frac{\sigma_{l,j,m}}{\sigma_{j,j,m}}\left(\mu_{k,m} - \frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\mu_{j,m}\right) + \frac{\sigma_{k,j,m}}{\sigma_{j,j,m}}\left(\mu_{l,m} - \frac{\sigma_{l,j,m}}{\sigma_{j,j,m}}\mu_{j,m}\right)\right)\mathbf{w}_m^{k\,T}\boldsymbol{\Phi}_m\mathbf{w}_m^l, \tag{3.15b}$$

and $c_{j,m}$ includes terms that do not have $X_{j,m}$ so there is no need to explicitly spell it out. It is clear that $Q_{j,m}$ and $L_{j,m}$ indicate the quadratic and linear contribution of each imaging feature $X_{j,m}$ to the classification accuracy, respectively.

***LSVM inverse operator***

The LSVM model takes the following form:

$$f(\mathbf{z}) = \mathbf{s}^{\boldsymbol{T}}\mathbf{z} + s_0, \tag{3.16}$$

where $\mathbf{s}$ and $s_0$ are estimated from the objective function $\min_{\mathbf{s},s_0,\boldsymbol{\xi}} \frac{1}{2}\mathbf{s}^{\boldsymbol{T}}\mathbf{s} + C\sum_i \xi_i$ subject to $y_i f(\mathbf{z}_i) \geq 1 - \xi_i$ and $\xi_i \geq 0\ \forall i$, where $C$ is the penalty parameter, $\xi_i$ is the slack variable for sample $i$ in a training dataset, $y_i$ is the class of sample $i$, and $f(\mathbf{z}_i)$ is the predicted value of sample $i$. Letting $\mathbf{s} = [\mathbf{s}_1^T, \ldots, \mathbf{s}_M^T]^T$, where $\mathbf{s}_m$ are the coefficients corresponding to $\mathbf{z}_m$, and substituting (3.6) into (3.16), the following is obtained

$$f(\mathbf{X}) = \sum_{m=1}^{M}\sum_{j=1}^{n_m} \mathbf{s}_m^T \mathbf{w}_m^j X_{j,m} + s_0. \tag{3.17}$$

It is clear from (3.17) that the magnitude of $\mathbf{s}_m^T\mathbf{w}_m^j$ indicates the contribution of each imaging feature $X_{j,m}$ to the classification accuracy. The sign of $\mathbf{s}_m^T\mathbf{w}_m^j$ indicates the direction of the contribution.

## 3.4 Clinical Application: A Glioblastoma Study

### 3.4.1 Background

Glioblastoma (GBM) is one of the most deadly types of cancer, with a median patient survival rate of 14 months (Sottoriva et al. 2013). One of the greatest challenges in

treating GBM is determining the optimal treatment therapies for different regions of the tumor, since GBM exhibits a broad intra-tumoral genetic variability. Namely, each tumor consists of several genetically distinct clonal populations that may require different types of therapy (Ene and Fine 2011). There has been a lack of available localized biopsy information for different regions of a tumor, which has caused most groups to sample a non-localized biopsy and use it to infer a single genetic profile for the entire tumor (Brown et al. 2008, Gutman et al. 2013, Jain et al. 2014, Itakura et al . 2015, Yang et al. 2015, Pope et al. 2008, Tykocinski et al. 2012, Gupta et al. 2015, Ryoo et al. 2013, Aghi et al. 2005). The latter approach may not be effective for treatment since the genetic profile of one region may not be characteristic of the genetic profile of another region, resulting in an incomplete or inferior treatment response (Marusyk et al. 2012). In essence, effective treatment of GBM needs a higher precision that goes beyond inter-tumor genetic difference but looks deeper into each tumor to characterize intra-tumor regional genetic variability.

To achieve this deeper level of precision, biopsy is a gold standard approach. However, biopsy is invasive, so that it is clinically infeasible to take a sufficiently large number of biopsy samples from each tumor in order to capture the regionally varying genetic landscape. On the other hand, magnetic resonance imaging (MRI) is non-invasive, enables assessment of the tumor in its entirety, and has shown capabilities of conveying a wide range of tumoral phenotypes that can potentially serve as surrogate markers for underlying genetics (Itakura et al. 2015, Stadlbauer et al. 2006, Hu et al. 2012a, Drabycz et al. 2010). Textural analysis of MRI images has been shown useful in characterizing the tissue structures in local areas of the image (Brown et al. 2008, Drabycz et al. 2010).

Building machine learning models that use localized image texture features to inform regional tumor genetics falls into the general research area of "radiomics", but it should be more accurately called "precision radiomics" due to its objective of deeply characterizing intra-tumor regional genetic heterogeneity.

### 3.4.2 Subject Selection and Image Acquisition and Preprocessing

*Patient recruitment:* Patients with clinically suspected with GBM and undergoing preoperative stereotactic MRI for surgical resection were recruited from Barrow Neurological Institute. It was confirmed that there was no previous treatment (including steroid administration), and approval was obtained from the institutional review boards. Written and informed consent was obtained from each subject prior to enrollment.

*Copy Number Variant (CNV) aberrations of interest:* The Cancer Genome Atlas (TCGA) has identified a set of biologically significant and highly recurrent DNA gains/losses through copy number analysis (Sottoriva et al. 2013, Brennan et al. 2013). These CNVs constitute known therapeutic targets and/or core GBM pathways; namely, RTK, PI3K, MAPK, p53, and Rb1 (Sottoriva et al. 2013, Brennan et al. 2013). For this study, tumor samples that demonstrated aberrations for each CNV were determined. To adequately power the radiogenomic models, CNVs were only included if they had alterations of at least 20% of the collection of tumor samples. From the biopsies that demonstrated sufficient abberations for a CNV, I built classification models to predict the status of each CNV (abberant vs. diploid/normal).

*Multiparametric MRI and ROI Segmentation:* Six multiparametric images were included in the present study, including T1+C, T2W, EPI+C, MD, FA, and rCBV (detailed

MRI protocols and image co-registration can be found in Hu et al. 2015 and the supplementary information). The T2W ROI of each tumor was manually segmented by a board-certified neuroradiologist.

*Texture analysis, image processing, and Principal Component Analysis (PCA):* Following image coregistration, all MRI data had uniform voxel size (1.2x1.2x3mm) across all the MRI contrasts (x,y,z dimensions). Regions of interest (ROIs) measuring 8x8x1 voxels (9.6x9.6x3mm) were generated at locations that correspond to each biopsy site. To ensure accuracy, a board-certified neuroradiologist visually inspected all ROIs. Before texture analysis, first order statistics were first acquired from raw image signals: mean (M) and standard deviation (SD) of gray-level intensities. Intensity values were then mapped within each ROI onto the range 0–255. This step helped standardize intensities between ROIs and reduced intensity non-uniformity effects on features extracted in subsequent texture analysis. Next, texture analysis was performed, incorporating 3 separate but complementary texture algorithms (as previously described (Brown et al. 2008, Tykocinksi et al. 2012, Drabycz et al. 2010, Haralick and Shanmugam 1973)): Gray Level Co-Occurrence Matrix (GLCM) (Urish et al. 2013), Local Binary Patterns (LBP) (Haralick and Shanmugam 1973), and Gabor filters (Grigorescu et al. 2002). 35 texture features were generated for each of six total MRI contrasts, which yielded 210 MRI-texture features and 12 raw features (i.e., mean and SD for six MRI contrasts) for a total of 222 image-based features for each ROI. Due to the high-dimensionality of image features relative to the sample size and the fact that features produced from the same algorithm and same contrast may be highly correlated, I performed PCA and determined Principal components (PCs)

for each texture algorithm-contrast combination – a total of 18 sets of PCA (Hu et al. 2015). The PCs were used in subsequent predictive modeling.

*Radiogenomic model using MMI-DDS:* I identified the subset of image-based PCs (determined from PCA above) with the greatest leave-one-out-cross-validated (LOOCV) area under the curve (AUC) for predicting the CNV status of each gene. LOOCV was used to avoid overfitting. AUC is a more robust metric than overall accuracy for a classifier due to its insensitivity to class imbalance. In the dataset, several genes are heavily imbalanced. To represent several types of classification methodologies, I separately applied three commonly used classification algorithms: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Support Vector Machines (LSVM) (Lin et al 2010, Dixon and Brereton 2009, Zacharaki et al. 2009). In building a classification model with sequential forward selection, the PC with greatest LOOCV AUC was first added to the model. A second PC was added whose incremental gain in LOOCV AUC is the largest among all the remaining PCs and the gain is statistically significant by a Hanley and McNeil's test (Hanley and McNiel 1983). Otherwise, only the first added PC is used in the classification model. This process continues with more PCs added if needed and only if the added PC improves the LOOCV AUC with statistical significance. Different classifiers (i.e., LDA, QDA, and LSVM) achieved the best performance in different genes. Therefore, I will report the results of the best classifier for each gene.

### 3.4.3 Results

*Patient, tissue samples, CNV aberrations of interest*: A total of 61 tissue specimens (21 BAT, 40 ENH) were collected from 18 patients. Of the core GBM pathways reported

by TCGA (Sottoriva et al. 2009, Brennan et al. 2013, Bonavia et al. 2011) CNVs associated with six driver genes met inclusion criteria for further radiogenomic analysis. *PTEN* deletion on 10q23 was the most commonly observed genetic alteration (80% of total samples), followed by *CDKN2A* deletion on 9p21.3 (72%), *RB1* deletion on 13q14 (59%), *EGFR* amplification on 7p11 (41%), *TP53* deletion on 17p13 (33%), and *PDGFRA* amplification on 4q12 (20%).

*Performance of MMI-DDS*: Table 4 shows the LOOCV AUCs for each gene. The significant PCs selected by MMI-DDS to be included in each model are provided in Table 5.

Table 4: PCs selected by SFFS to include in the radiogenomic models from conventional and advanced MRI sequences T1W+C, T2W, EPI+C, rCBV, FA, and MD. The PC number from each contrast-texture algorithm combination is shown in parentheses. The p-value of each additional PC beyond the first one that was added to the model is also shown.

| CNV | PCs selected by SFFS<br>All |
|---|---|
| **EGFR ++ (7p11)** | MD-GLCM (1) |
| **PDGFRA ++ (4q12)** | MD-GLCM (2) |
| **PTEN – (10q23)** | FA-LBP (5) |
| **CDKN2A – (9p21.3)** | T2W-LBP (2)<br>T1W+C-LBP (1); p<0.03 |
| **RB1 – (13q14)** | FA-LBP (2)<br>T1W+C-LBP (1); p<0.05 |
| **TP53 – (17p13)** | rCBV-LBP (5) |

Table 5: LOOCV AUCs on all samples, BAT samples, and ENH samples for conventional and advanced MRI sequences T1W+C, T2W, EPI+C, rCBV, FA, and MD.

| CNV | LOOCV AUC | | |
|---|---|---|---|
| | Overall | BAT | ENH |
| **EGFR ++ (7p11)** | 0.70 | 0.72 | 0.69 |
| **PDGFRA ++ (4q12)** | 0.77 | 0.85 | 0.75 |
| **PTEN – (10q23)** | 0.71 | 0.44 | 0.92 |
| **CDKN2A – (9p21.3)** | 0.81 | 0.82 | 0.80 |
| **RB1 – (13q14)** | 0.78 | 0.72 | 0.81 |
| **TP53 – (17p13)** | 0.68 | 0.61 | 0.70 |

### 3.4.4 Discussion

GBM's intratumoral genetic heterogeneity, hypothesized to derive from clonal expansion of multiple genetically divergent tumor populations, requires targeted therapy to mitigate tumoral resistance. A clonal population may express varying sensitivities and drug targets, which increases the chance that pre-existing resistant clones will result in failed treatment therapy and subsequent tumor recurrence. Adjacent clonal populations can also exert influence on therapeutic response through biological interactions (Ene and Fine 2011, Marusyk et al. 2012, Bonavia et al. 2011). Thus efforts are needed to develop combinatorial strategies that can take advantage of genetic heterogeneity to enhance current therapeutic methods (Ene and Fine 2011, Marusyk et al. 2012, Bonavia et al. 2011). As genetically informed technologies become more feasible, characterizing intratumoral heterogeneity will have an even greater role in strategizing effective targeted therapies.

Although CE-MRI is used to help neurosurgeons when collecting surgical biopsies from the ENH, CE-MRI alone lacks the precision to predict regional, genetically distinct clonal sub-populations within each tumor. As a result, other imaging features have been evaluated as potential biomarkers for genetic status (Brown et al. 2008, Gutman et al. 2013, Jain et al. 2014, Itakura et al. 2015, Yang et al. 2015, Pope et al. 2008, Tykocinski et al. 2012, Gupta et al. 2015, Ryoo et al. 2013, Aghi et al. 2005, Barajas et al. 2010). However, most of these studies fail to be informative of intratumoral heterogeneity since they used non-localizing biopsies (usually from a small representative sub-region) to infer a single genetic profile for an entire tumor. This method is sub-optimal since genetic profiles from one biopsy location may not accurately correspond with those from other tumor sub-regions. Gutman et al. 2013 and Jain et al. 2014 independently reported a lack of correlation between imaging features and common GBM drivers such as EGFR, PDGFRA, PTEN, and CDKN2A. These drivers typically show regional intratumoral heterogeneity (Sottoriva et al. 2013, Van Meter et al. 2006). Other studies have reported imaging that has mixed correlations with GBM subtypes (Jain et al. 2014, Yang et al. 2015). However, these studies did not take into consideration that multiple subtypes can exist together in a single tumor (Sottoriva et al. 2013). Additionally, several groups using non-localizing biopsies in their analysis have conflicting results on whether perfusion MRI measures correlate with EGFR (Jain et al. 2014, Tykocinski et al. 2012, Gupta et al. 2015, Ryoo et al. 2013, Aghi et al. 2005).

Other studies have also utilized image-guided biopsies (Barajas et al. 2010, Van Meter et al. 2006). However, my study is different in that it includes development of

clinically interpretable models for driver genes known to play a role in GBM. These facets have facilitated the recognition of several significant associations between regional CNV status and imaging features.

I used classification algorithms (i.e., LDA, QDA, and SVM) coupled with sequential forward selection (SFFS) to identify subsets of image-based PCs that achieved the highest LOOCV AUC for each gene. The model development presented in this paper is limited and can be augmented in the future with additional models such as regression, artificial neural networks, Bayesian networks, and deep learning (if the dataset being analyzed is large enough).

Additional limitations to the analysis of the current analysis are as follows: (1) Since this study examines a small data set, the derived models need to be validated in a larger GBM cohort. Having a larger dataset should also increase the ability to capture more GBM driver gene alterations (e.g., CDK4, c-MET, etc.), which were too infrequent in the current cohort to sufficiently characterize through imaging. Prospective validation can also aid targeting of biopsies for genetically diverse regions within each tumor, which can facilitate integration these predictive models with surgical neuronavigation. (2) Misregistration errors may be present because of image distortions as well as brain shift post craniotomy. To minimize these errors, small craniotomy sizes are taken to reduce brain shift and stereotactic image location were visually validated with intracranial neuroanatomic landmarks to help adjust for random brain shifts. Potential geometric distortions were also reduced by rigid-body coregistration of stereotactic and DSC-MR imaging (Hu et al. 2015, Hu et al. 2012a, Barajas et al. 2010, Barajas et al. 2012, Hu et al.

2012b). Combined misregistration is estimated to be about 1–2 mm from both brain shift and registration technique—similar to that from previous studies by using stereotactic needle biopsy (Stadlbauer et al. 2006). Additionally, multiple tissue samples from spatially distinct subregions were collected within the same tumor for each patient. To minimize potential effects of sample overlap, small ROI sizes were used. So impact from these minority samples is estimated to be negligible.

## 3.5 Clinical Application: A Migraine Study

Approximately 36 million Americans suffer from migraine (Daniel and Mauskop 2016). Current clinical diagnosis is primarily symptom-based, which is prone to patient subjectivity. Imaging has shown great promise for providing objective measures of the disease and for improving the diagnostic accuracy (Schwedt et al. 2015, Chong et al. 2017). However, most existing research on migraine diagnosis focuses on single modalities. In this section, I present a study of using MMI-DDS to integrate multi-modality structural and functional imaging data for migraine diagnosis.

### 3.5.1 Subject Selection and Image Acquisition and Preprocessing

The data used for this application were obtained from Mayo Clinic Arizona and Washington University School of Medicine in St. Louis: A total of 106 subjects who had structural and functional MRI data were included in this analysis, consisting of 57 individuals with migraine (PMs) and 49 healthy controls (HCs). These 106 subjects were a subset of subjects included in prior analyses (Schwedt et al. 2015, Chong et al. 2017).

PMs were diagnosed in accordance with the diagnostic criteria defined by the International Classification of Headache Disorders (Arnold 2018)

Structural MRI data were obtained from two Siemens 3T MRI machines. Using a cortical reconstruction and segmentation program in the FreeSurfer image analysis suite (version 5.3, http://www.surfer.nmr.mgh.harvard.edu/), cortical area, thickness and volume measurements of 68 ROIs were extracted. Additionally, resting-state functional connectivities, i.e., fMRI data, were collected for each subject. Standard Statistical Parametric Mapping (SPM) methods were used to preprocess the fMRI data. Specifically, fMRI signals were temporally filtered between 0.01 to 0.1 Hz to retain the low frequency components. Variance relating to signals of no interest was removed through linear regression. 33 ROIs were chosen based on commonly cited regions for which PMs show abnormalities (Mainero et al. 2011, Russo et al. 2012). Among the 33 ROIs, there are 16 pairs; each pair consists of two regions with the same name but located at the left and right sides of the brain, respectively. The remaining one ROI is located in the middle of the brain. Each pair of ROIs was aggregated into one ROI by averaging their respective time courses. This reduces the number of ROIs to 16+1=17. Partial correlations between the 17 ROIs were computed, forming 136 connectivity features. Note that I also tried keeping the original 33 ROIs without pair-wise aggregation, but the result was not as good as the one with aggregation.

In summary, this study utilizes two imaging modalities in terms of the image acquisition techniques, i.e., structural MRI and fMRI. Structural MRI produces three sets of features for 68 ROIs, i.e., area features, thickness features, and volume features. Because these three sets measure different aspects of the brain structure, they are treated as three

modalities in my analysis. As a result, four modalities are used in MMS-DDS, including cortical area (68 features), thickness (68 features), volume (68 features), and resting-state functional connectivity (136 features).

### 3.5.2 Classification Accuracy by Multi-Modality Imaging Data Integration

In this experiment, I show the performance of my system in integrating all the imaging modalities. Specifically, I first apply modality-wise PCA to each modality and keep the PCs that explain 85% of the variance in the data of the respective modality. Then, cPSO takes as input the data on the combined PC set across all the modalities. The optimal parameter $K$ for cPSO is found to be $K^* = 8$, 6, and 9, respectively. $K$* was chosen as the value at the "elbow" of the plot of CV errors against different values of $K$. Table 4 (last column) shows the CV classification errors corresponding to LDA, QDA, and LSVM under their respective $K^*$. For comparison, I also apply my system to integrating the three sets of features from structural MRI, i.e., cortical area, thickness, and volume, and the result is shown in the first column of Table 6. Furthermore, I report the result on using resting-state functional connectivity from fMRI alone. These analyses aim to show the benefit of integrating structural and functional imaging data.

Table 6: CV classification errors (avg ± std error) of the proposed MMI-DDS applied to MRI alone, fMRI alone, and MRI+fMRI combined

| | MRI (area+thickness+volume) | fMRI | MRI+fMRI |
|---|---|---|---|
| **LDA** | 24.43% ± 0.79% | 27.17% ± 0.74% | 21.79% ± 0.50% |
| **QDA** | 26.32% ± 0.53% | 29.72% ± 0.75% | 22.45% ± 0.48% |
| **LSVM** | 20.38% ± 0.63% | 25.38% ± 0.89% | 17.17% ± 0.19% |

In all three classifiers, the system's ability for integrating data from structural and functional imaging modalities is evident. Using a two-sample t-test, the CV error of MRI+fMRI is significantly lower than MRI alone with p values of 0.0062, $2.2 \times 10^{-5}$ and $2.8 \times 10^{-4}$ for LDA, QDA, and LSVM, respectively. Because the CV errors of MRI are lower than fMRI, there is no need to compare MRI+fMRI with fMRI. I conclude the integration of multi-modality imaging can significantly improve the diagnosis accuracy. Furthermore, among the three classifiers, LSVM achieves the lowest error, i.e., highest accuracy of 83%, using MRI+fMRI.

Please note in the single modality migraine study (Schwedt et al. 2015) where structural MR data were analyzed, the classification accuracy was 68%; and the single modality migraine study using fMRI data had 81% classification accuracy (Chong et al. 2017). One may argue that the 83% accuracy reported in this study is a marginal improvement compared to 81% accuracy. I contend that first, Table 6 indicates the statistical differences between the two approaches (fMRI+MRI vs. fMRI) using the same features sets. Second, a voxel-by-voxel connectivity approach was adopted in (Chong et al. 2017) while 136 features measuring the correlations among 17 ROIs were used in this research. Since one of the key traits of the proposed MMI-DDS is *interpretability*, the use of a ROI based approach may have easy adoption in clinical practice. It is certainly the interest of the team to explore the use of a voxel-by-voxel approach to investigate whether a better accuracy may be achieved from this dataset.

### 3.5.3 Biomarker Identification

For each classification model in the last column of Table 4, I apply the proposed clinical utility engine to find the contribution of each feature in the respective imaging modality. Because LSVM gives the highest accuracy, next I examine the result for LSVM more closely. Specifically, I would like to focus on the features that have large positive or negative contributions to the classification accuracy, i.e., features whose contribution weights are large in magnitude. These features have higher likelihood of being potential migraine biomarkers. To this end, I pool the weights from all the modalities together and rank them from the largest to the smallest in terms of their magnitudes. This would give us a rank for the features. Table 7 lists the features that rank in the top 5%. These roughly correspond to features that are significant at 0.05 significance level, a common choice for assessing statistical significance. Figure 5 highlights the ROIs corresponding to the area features in Table 7 on the brain surface. Figure 6 shows the resting-state functional connectivity in Table 7 on the brain surface.

Table 7: Imaging features that rank in the top 5% in terms of the magnitudes of contribution weights for LSVM (L: left hemisphere of the brain; R: right hemisphere of the brain)

| Feature set | Features |
|---|---|
| Area (MRI) | Frontal pole (L), Inferior temporal (L), Middle temporal (L), Transverse temporal (L), Transverse temporal (R), Banks of the superior temporal (R), Precentral (R), Paracentral (R), Entorhinal (R) |
| Thickness (MRI) | Insula (R) |
| Volume (MRI) | None |
| Resting-state functional Connectivity (fMRI) | <Posterior cingulate, Dorsolateral prefrontal><br><Anterior cingulate, Amygdala><br><Inferior lateral parietal, Supplementary motor><br><Primary somatosensory, Temporal pole><br><Temporal pole, Caudate><br><Middle cingulate, Secondary somatosensory><br><Inferior lateral parietal, Temporal pole> |

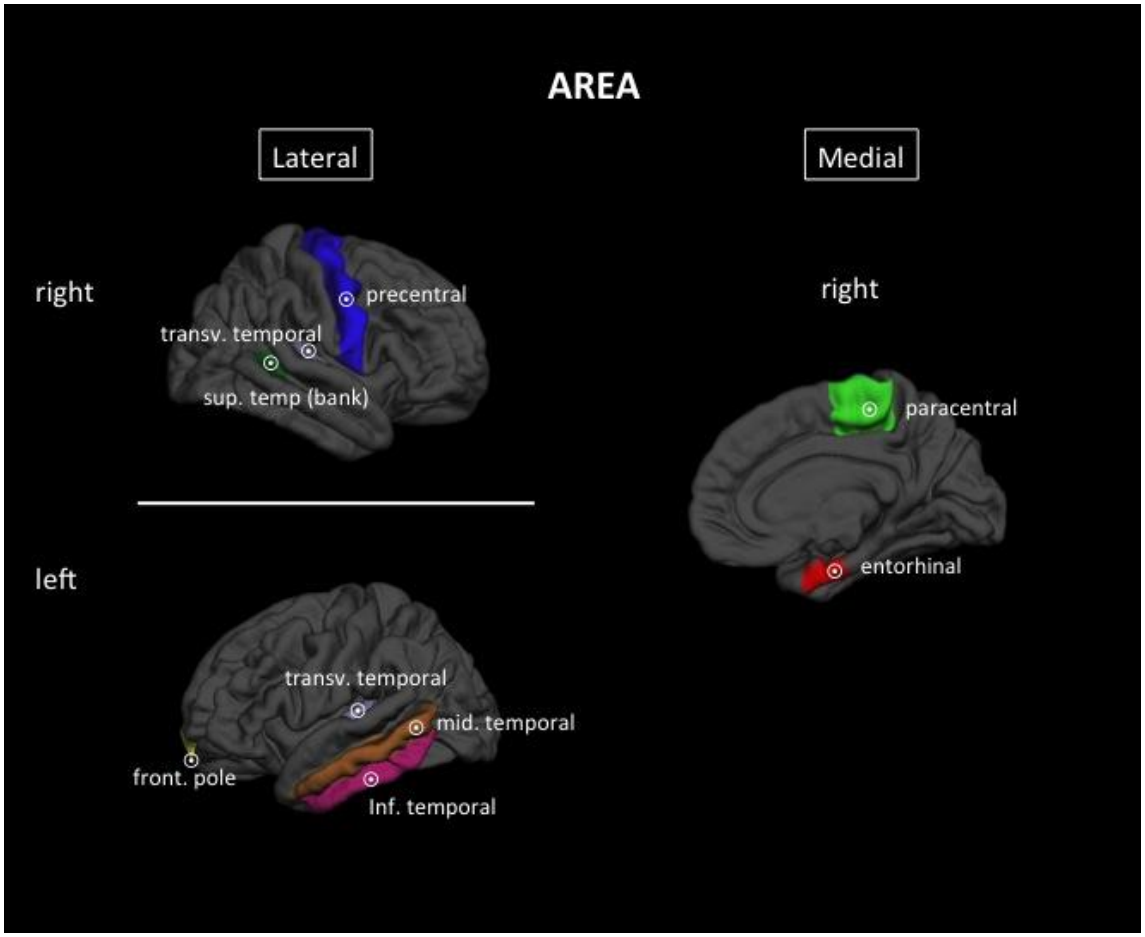



Figure 5: ROIs corresponding to the area features in Table 7 shown on an average inflated brain surface. front. pole=frontal pole; inf. temporal=inferior temporal; mid. temporal=middle temporal; sup. temp (bank)= bank of the superior temporal; transv. temporal=transverse temporal

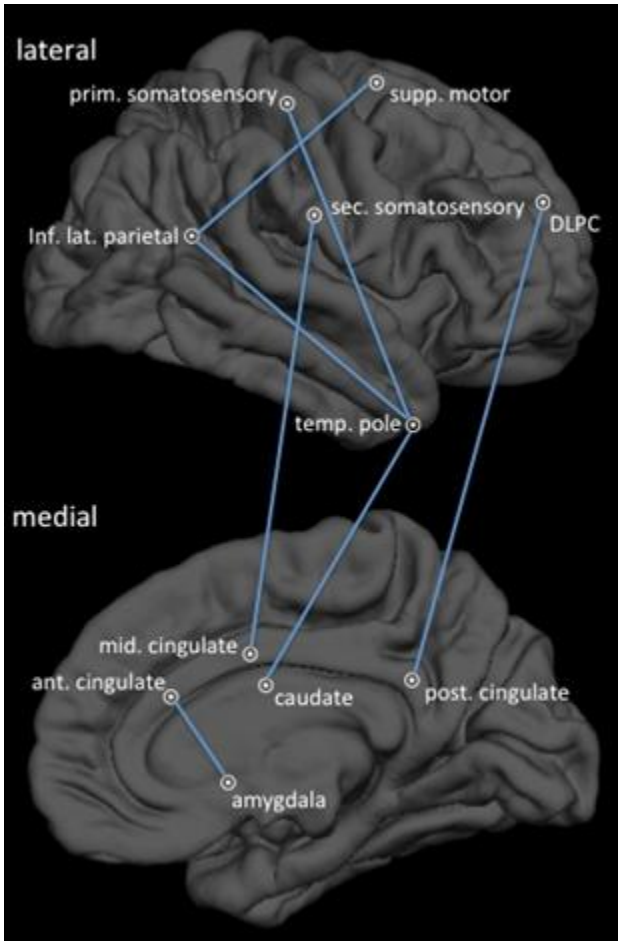



Figure 6: Resting-state functional connectivities corresponding to Table 7. For illustration purposes, functional connectivities are shown on an inflated right hemisphere average brain surface. DLPC=dorsolateral prefrontal; ant. cingulate=anterior cingulate; inf. lat. parietal=inferior lateral parietal; mid. cingulate=middle cingulate; post. cingulate=posterior cingulate; prim. somatosensory=primary somatosensory; sec. somatosensory=secondary somatosensory; supp. motor=supplementary motor; temp. pole=temporal pole

As expected, given the symptoms of migraine, brain regions most contributing to migraine classification (those listed in Table 7) play important roles in pain processing and processing of multisensory stimuli. Whereas some are regions that are predominantly responsible for sensory-discriminative pain processing (e.g. somatosensory cortex), others are responsible for affective-emotional processing (e.g. amygdala, anterior cingulate

cortex), cognitive processing (e.g. prefrontal cortex), or integration of incoming sensory information from different domains (e.g. temporal pole). Several of these regions have commonly been identified as having atypical structure or function in previous migraine studies. The temporal pole, a multisensory region that integrates somatosensory, visual, auditory, and olfactory stimuli (Schwedt 2013), has frequently been identified to have atypical structure, function and functional connectivity in migraine studies (Schwedt et al. 2014, Rocca et al. 2006). Atypical function of the temporal pole in PMs might contribute to common migraine symptoms such as the exacerbation of migraine headache intensity when exposed to lights and sounds. The anterior cingulate cortex is involved in affective components of pain processing including pain anticipation (Palermo et al. 2015), and has been shown to have atypical activation, structure, and functional connectivity in PMs (Russo et al. 2012, Jin et al. 2013, Schwedt et al. 2013). The amygdala and middle cingulate cortex are also involved with determining pain affect, with the middle cingulate cortex possibly having additional roles in the integration of other aspects of pain processing (e.g. sensory discriminative, affective, cognitive) (Palermo et al. 2015, Simons et al. 2014). One fMRI study on heat pain processing found that interictal PMs showed stronger middle cingulate cortex activation than HCs (Schwedt et al. 2014). PMs have also been demonstrated to have atypical stimulus-induced activation of the amygdala during migraine attacks and atypical functional connectivity of the amygdala compared to HCs (Schwedt et al. 2013, Stankewitz and May 2011). My findings are consistent with these previous findings.

## 3.6 Conclusion

In this chapter, I developed a clinical decision support system, MMI-DDS, that integrates multi-modality imaging data for disease diagnosis. The system was designed to achieve flexibility, sufficient accuracy, and interpretability, which are three important traits required for clinical decision support systems, but unfortunately are inadequately addressed by prior research. Specifically, my proposed system included a modality-wise PCA, a cPSO algorithm for classification, and a clinical utility engine for identifying contributing features to facilitate biomarker identification. I applied the proposed MMI-DDS to using multiparametric MRI to predict intra-tumor genetic variability of glioblastoma brain cancer. A high AUC of 0.81 was achieved for predicting the CDKN2A-aberrant CNV. I also applied MMI-DDS to migraine diagnosis by integrating cortical thickness, area, and volume data acquired from structural MRI and resting-state functional connectivity data from fMRI. A high accuracy of 83% was achieved by integrating the structural and functional modalities together, which is significantly better than using single modalities alone. Furthermore, the clinical utility engine identified contributing features to the classification accuracy. Highly ranked features according to their respective contributing weights were found to be relevant to migraine as confirmed by existing studies. Future research includes extending the system's capability to multi-class classification that is useful for disease subtype classification, and to prediction of numerical response variables such as disease severity.

# CHAPTER 4

# INTEGRATED FEATURE AND INSTANCE SELECTION IN SEMI-SUPERVISED REGRESSION FOR SMARTPHONE-BASED TELEMONITORING OF PARKINSON'S DISEASE PATIENTS

## 4.1 Background

As mobile phone technology has improved in recent years, there has been an unprecedented opportunity to collect high-resolution data from users. Now that approximately 77% of American adults own a smartphone (according to 2018 Pew Research surveys), there is capability to collect this high-resolution data on a large scale. Because smartphones are equipped with many sensors, they are capable of collecting an abundance of useful information on user's daily activities, including data measured by the microphone, camera, accelerometer, and gyroscopes. With this large amount of information, there is an increasing number of endeavors to improve healthcare through mobile-powered patient portals, mobile health (mHealth) apps, and telemedicine. There have been recent efforts to utilize this technology for telemonitoring of Parkinson's Disease (PD). PD is the second most common neurodegenerative disease and affects seven to ten million people worldwide (Goetz et al. 2009). PD is a movement disorder characterized by a lack of dopamine production in cells of the midbrain. Common symptoms include speech changes, voice tremor, slowed movement, tight muscles, and loss of balance. Although there is no known cure for PD, effective treatment can slow down and ameliorate the progression of the disease.

In order to have effective treatment, the disease progression and severity must be monitored on a frequent basis. Typically, this requires the patient's presence in a

specialized clinic, which is expensive and burdensome to both the patient and medical staff. Thus, having inconsistent evaluations in disease condition is not an uncommon occurrence, which results in insufficient therapy since the treatment will always be behind the disease progression.

Smartphones have emerged as an alternative to provide an inexpensive and consistent way of monitoring symptoms of PD. Activity collected from smartphones can be transformed to useful features that can help better understand characteristics of the disease and monitor disease progression. Sage Bionetworks created a mobile application called mPower for smartphone telemonitoring of PD patients (http://sagebionetworks.org/research-projects/mpower-researcher-portal/). mPower collects several different types of information by having the user perform activities such as speaking, walking, memorizing, and tapping. This information is sufficient to measure the different symptoms of PD with minimal interruption to the patient's daily routine since these activities can be performed at home. Information can be generated from these exercises in the form of features that can be used for better inferring disease progression. Because using mPower is convenient and low-cost, the patient's PD status and progression can be assessed consistently and therapy can be adjusted in a timely manner to provide the best treatment.

In order to utilize features generated from the mPower app, statistical modelling techniques can be used to decipher the features into some indicator of disease severity. To quantify disease severity, the most common metric is the Unified Parkinson's Disease Rating Scale (UPDRS). The UPDRS is generated as a summary score from 42 question-survey administered to the patient to assess PD-related symptoms. Developing a model

that connects the UPDRS score to features generated from the tasks on the mPower app, would provide huge value to monitoring patient disease condition on a consistent basis.

However, predictive modeling of PD using telemonitoring signals of PD patients has the following issues: First, there is an insufficient number of instances in which the patient visited the clinic to obtain a UPDRS score and has corresponding telemonitoring signals from their smartphone. This provides inadequate information from which to build an accurate model. Semi-supervised learning methods are needed to also utilize instances that do not have a corresponding UPDRS score to build an accurate model. Second, there is plethora of features collected from smartphones from which to train a model, however many of the features are irrelevant to predicting UPDRS. Thus feature selection techniques must be employed to determine the optimal subset of available features to include in the final model.

To address these challenges, I introduce a first-of-its-kind semi-supervised feature selection algorithm for continuous prediction of Parkinson's Disease severity. This approach combines particle swarm optimization (PSO) for selection of smartphone-based telemonitoring features and semi-supervised learning (SSL) to utilize all available data collected from smartphones. I further extend this algorithm by introducing a graph sampling method that reduces the computational time and trains the model on a smaller-representative subset of the larger training data population. Because the proposed model integrates both feature and sample selection with SSL, it is named s2SSL, implying an SSL with selection in two aspects: feature and sample. s2SSL aims to balance data inclusivity (through SSL) and usability (through feature and sample selection).

The remainder of this chapter is organized as follows: Section 4.2 provides a literature review of semi-supervised learning, feature selection, and sampling techniques relevant to s2SSL. Section 4.3 presents the methodological development of s2SSL. Section 4.4 provides some simulation tests of s2SSL to demonstrate the utility of different aspects of the algorithm. Section 4.5 provides an application study of s2SSL used on smartphone-based telemonitoring of Parkinson's Disease patients. Section 4.6 discusses the results and concludes the chapter.

## 4.2 Literature Review

### 4.2.1 Semi-supervised learning

Semi-supervised learning (SSL) has been widely used in applications in which labeled data are scarce but unlabeled data are available in large quantity. There are many types of SSL algorithms, including generative, self-training, co-training, low-density separation, and graph-based models.

Supervised generative models assume that the data take on a probability distribution $p(\mathbf{x}, y) = p(y)\ p(\mathbf{x}|y)$, where $\mathbf{x}$ are features, $y$ is the response variable, and $p(\mathbf{x}|y)$ follows some identifiable mixture distribution. Semi-supervised generative models incorporate the probability distribution of unlabeled data such that the probability distribution becomes $p(\mathbf{D}) = p(y)\ p(\mathbf{x}_L|y)p(\mathbf{x}_U)$, where $\mathbf{D} = \{(\mathbf{x}_L, y), \mathbf{x}_U\}$ is the dataset consisting of labeled $(\mathbf{x}_L, y)$ and unlabeled $(\mathbf{x}_U)$ samples. Using an optimization method such as expectation maximization (EM), parameters of the mixture distribution can be estimated. For example, Holub et al. used a generative model to transform face image data into fixed-length Fisher

score vectors, and inputted the transformed features to a standard discriminative classifier (Holub et al. 2005). Fujino et al. introduced a hybrid generative-'bias correction' model for text classification based on the maximum entropy principle (Fujino et al. 2005).

Self-training starts by training a classifier using only labeled data, then uses the classifier to predict unlabeled data and adds most confidently predicted unlabeled samples to re-train the classifier. Li et al. presented self-trained support vector machines (SVMs) and applied them to an electroencephalography (EEG)-based brain computer interface (BCI) speller (Li et al. 2008). Tanha et al. modified the basic decision tree learner for self-training and applied the model to several datasets from the UCI Machine Learning Repository (Tanha et al. 2017, Bach and Lichman 2013).

Co-training extends the idea of self-training by training separate classifiers on two sub-feature sets and adding most confidently predicted unlabeled samples by one classifier to the other classifier's re-training process. Wan used co-training of a sentiment classifier to utilize abundant information of English sentiment classification and unlabeled Chinese data for the problem of cross-lingual sentiment classification (Wan 2009). Reference Zhou et al. showed that by utilizing the correlations between the two feature subsets using canonical correlation analysis, co-training can be successfully performed using only one labeled training sample (Zhou et al. 2007).

Low-density separation aims to find a classification boundary that separates not only labeled data of different classes but also unlabeled data at a low-density region in the feature space. Reference Zhu and Lafferty demonstrated a heuristic to minimize the average label entropy by utilizing a harmonic function summed over unlabeled data (Zhu and Lafferty 2005). Lawrence and Jordan proposed a Gaussian Process with a "null

category noise model" and demonstrated its performance on labeled and unlabeled handwritten digits (Lawrence and Jordan 2005).

Graph-based SSL has recently become popular because of its relatively high accuracy and efficiency. The basic idea is to construct a graph with vertices being labeled and unlabeled samples in a training set and edges weighted by vertex proximity in the feature space. There are two types of graph-based SSL: transductive and inductive learning models. The former aims to formulate a method to propagate label information from labeled samples to unlabeled samples in a specific dataset. In this way, the unlabeled samples in the dataset are classified/predicted. The latter aims to train a model using labeled and unlabeled samples, which is not only used to predict the unlabeled samples in training but also new samples.

For transductive learning, Zhu et al. 2003 proposed a Gaussian random field model with the mean of the field characterized in terms of harmonic functions. They tested the model on digit and text classification tasks. Zhou et al. 2004 introduced a local and global consistency framework based on the quadratic loss of prediction on labeled samples regularized by a normalized Laplacian matrix. For inductive learning, Zhu and Lafferty 2005 regularized generative mixture models with graph Laplacian and demonstrated its performance on handwritten digit and teapots image datasets. Belkin et al. 2006 proposed the manifold regularization (MR) framework, which relied on properties of reproducing kernel Hilbert spaces (RKHS) to enable efficient and accurate prediction.

### 4.2.2 Semi-supervised Feature Selection

Just as in conventional feature selection, semi-supervised feature selection can be divided into three areas—filter method, wrapper method, and embedded method (Sheikhpour et al. 2017).

Filter methods are known to be very efficient in selecting features as they are primarily a screening technique that reduces features before a model is trained. They examine the inherent properties of the labeled and unlabeled data to choose features prior to training a model. Laplacian score ranks features based on their locality preserving power, i.e., features that preserve the underlying geometry of the data (Cheng et al. 2011, Zhao et al. 2008, Doquire and Verleysen 2013). The Fisher criterion chooses features based on their discriminant capability (their ability to minimize within-class variance and maximize between-class distance), and is the SSL analogue of the classic Fisher score (Chen et al. 2010, Yang et al. 2010, Yang et al. 2011, Liu et al. 2013, Liu et al. 2010). RELIEF-based methods are also emerging as an effective method for calculating a weight vector that ranks features based on the differences of given samples and their nearest neighbors. Currently there are RELIEF methods that can handle two-class and multi-class data (Cheng et al. 2008, Tang and Zhang 2018). Other filter methods include those based on pairwise constraints (Kalakech et al. 2011, Benabdeslem and Hindawi 2011), spectral graph theory and cluster assumption (Zhao and Liu 2007), sparse models (Han et al. 2015), local discriminative information (Zeng et al. 2016), and conditional mutual information/entropy (Quinzán et al. 2009). The disadvantage of using filter methods are that the feature selection is not integrated with the model training.

Wrapper methods utilize an existing learner (or an ensemble of learners) in a framework to choose the optimal number of features. Methods based on a single learner (Ren et al. 2008, Wang et al. 2008) use self-training based semi-supervised learning. Initially, a supervised learner is trained on the labeled instances, employing a greedy feature selection method such as sequential forward selection (SFS). The selected features are then used to train a model to predict the labels of the unlabeled data. Predictions are selected as pseudo-labels to augment the current labeled dataset (either through random selection or based on confidence in prediction), then a new feature selection model is trained. This process is repeated several times to accumulate several different subsets of selected features, and those with the highest frequency are chosen for the final model. Ensemble learner methods (Bellal et al. 2012, Han et al. 2011, Barkia et al. 2011) utilize multiple classifiers and then combine their output results either through self-training or co-training. A confidence measure is then used to select unlabeled data as pseudo-labels augment the labeled dataset. Different labeled training sets are typically created through methods like bagging, and different feature subsets are generated through random subspace methods (RSMs).

Embedded methods incorporate feature selection in the model training process utilizing labeled and unlabeled data. Methods based on support vector machines (Yang and Wang 2007, Xu et al. 2010, Ang et al. 2015, Dai et al. 2013) maximize the decision boundary margin between classes while incorporating the local structure of the labeled and unlabeled data instances. Embedded feature selection based on sparse models and the graph Laplacian (Song et al. 2016, Ma et al. 2012, Shi et al. 2014, Ma et al. 2011) employ a variety of sparse model techniques and graph-based semi-supervised learning to utilize

the information in the labeled and unlabeled instances. Manifold regularization (MR) is the most popular in this group, as it can use the graph Laplacian to diffuse information in the labeled instances to unlabeled instances in a way that can easily integrate with many existing algorithms (Ma et al. 2012, Ma et al. 2011).

In the existing literature to date, the only algorithm that is directly applicable to semi-supervised feature selection in regression problems (with a continuous response) is the Laplacian score (Doquire and Verleysen 2013). However, the graph Laplacian that is built to rank the features to be chosen is calculated from noisy and relevant features, which causes the features to be chosen based on a suboptimal criteria. Thus there is a need to create a semi-supervised feature selection method that selects a relevant subset based on a more proper criterion.

### 4.2.3 Graph sampling

Manifold regularization (MR) has become a popular technique for semi-supervised learning. However there are some limitations on larger datasets due to the computational complexity of incorporating a matrix-embedded graph into model training, and there is a need to develop a sampling technique to minimize computational time and train on the most relevant instances.

One category of graph sampling techniques is embedded directly into the objective function to be minimized. Studies by Zhang et al. and Zuo et al. incorporated the $L_1$-norm with respect to the kernel coefficients of the response for manifold regularization (MR) (Zhang et al. 2014, Zuo et al. 2015). By the classical Representer Theorem, I can find the

optimal solution for MR to be $f^*(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x)$, where $\alpha_i$ is the Lagrangian multiplier for instance $i$, and $K(x_i, x)$ is the Mercer kernel associated with instance $i$. To reduce the computation complexity of calculating this solution, an $L_1$-norm can be applied to the $\alpha_i$'s that cause some of the $\alpha_i$'s to shrink to zero.

Additionally, there are studies that employed a Laplacian $L_1$-norm into the objective function (Lu et al. 2015, Lu and Wang 2015). The intrinsic Laplacian regularization term in MR can be modified to become an $L_1$ regularizer by reducing the Laplacian matrix to be $L = V\Sigma^{-1}V^T = (\Sigma^{1/2}V^T)^T\Sigma^{1/2}V^T = B^TB$, where $V$ is the set of eigenvectors of $L$ and $\Sigma$ is a diagonal matrix of the eigenvalues for $L$. Once $B$ is derived, it can be used to substitute an $L_1$ version of $\mathbf{f^T}L\mathbf{f}$ (i.e., $|B\mathbf{f}|_1$) in the MR formula, which will induce sparsity based on the Laplacian.

Other sampling methods are used to reduce the population to a representative set before SSL model training. Performing sampling before model training can have a significant time advantage since the initial graph size before model training is much smaller. Wang and Zhang used a graph-based sampling method to eliminate bridge points, i.e., instances that are noisy or do not have many nearest neighbors in a given search radius (Wang and Zhang 2008). By performing eigendecomposition of the covariance matrix of instance $i$, one can calculate the confusion rate of each covariance matrix as $c_i = \sum_{k=1}^{\mathrm{K}} \lambda_{ik}/\bar{\lambda}_{ik}$, where $\lambda_{ik}$ is the $k^{\text{th}}$ eigenvalue of the covariance matrix for $\mathbf{x}_i$, $\bar{\lambda}_{ik} = \sum_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} \lambda_{jk}$, where $\mathcal{N}(\mathbf{x}_i)$ is the neighborhood of points around $\mathbf{x}_i$. Points with a higher confusion rate, $c_i$, have a higher chance of being bridge points. Typically, this procedure is performed by visually examining the graph, or by running several tests with different

values of confusion rate thresholds and choosing the threshold that maximizes the accuracy. For continuous $Y$, this procedure can be used to eliminate noise or outliers.

Block sparsity (Zhao et al. 2014) used the $L_{2,1}$-norm to make instances in the same class share the same sparse pattern. Using the framework of sparse congruency representation, the method first solves the problem $\min_{Z,E} \|Z^T\|_{2,1} + \|E\|_{2,1}$, subject to $X = XZ + E, Z \geq 0$, where $X_{d\times N}$ is the data matrix with $d$ dimensions and $N$ instances, $Z_{N\times N}$ is the reconstruction coefficients matrix, and $E_{d\times N}$ is the noise term. After solving for $Z$, one can derive a Laplacian matrix to be used for SSL.

Goldberg et al. 2009 introduced a cover sampling technique that selects one instance at a time and removes unlabeled data. Given a few labeled data and many unlabeled data, the algorithm retains all the labeled data and derives an approximate cover of the unlabeled data as follows: (1) choose a random unlabeled point, $x^{(0)}$, (2) remove its unlabeled neighbors $N(x^{(0)})$, (3) select $x^{(1)}$—the next nearest neighbor to $x^{(0)}$, and (4) repeat the procedure until there are no more points to eliminate. This procedure is attractive in that it significantly reduces the number of unlabeled instances, but it is greedy in the way that it creates the cover since it chooses each $x$ one-at-a-time.

Sun et al. 2014 employed a method that favors instances that have a higher degree in the underlying graph. This seems to be an intuitive approach since it retains instances that have the most connections, and thus play a more important role in the underlying manifold of the data. The paper makes an interesting proposal to reduce the instances such that the inherent manifold that underlies the instances is retained. The objective function that is proposed is

$$\max \frac{1}{m-t} \sum_{i=t+1}^{m} \left( \max_{j=1,\ldots,t} W_{ij} \right) \tag{4.1}$$

where $m$ is the number of vertices in the original graph $G$; $t$ is the number of vertices the user chooses to have in the reduced (sparse) graph $G_S$; and $W_{ij}$ is the weight between instance $i$ in $\bar{G}_S$ and instance $j$ in $G_S$. In short, this objective function attempts to find $G_S$ from $G$ such that the sum of the maximum weights between each node outside of $G_S$ (i.e., $\bar{G}_S$) and each node within $G_S$ is maximized. This objective encourages a high spatial connectivity between $G_S$ and $G$. Encouraging this high spatial connectivity has two functions—(1) to remove outliers/noise, and (2) to break up the domination of instances that are very close to each other.

The problem of obtaining a manifold preserving graph $G_S$ defined by (4.1)is NP-hard, so the authors propose a greedy method to solve it. In each of the *t* iterations, the greedy algorithm chooses the vertex with highest degree in $G$, adds it to $G_S$, removes all edges associated with that vertex, and repeats the process until the sparse graph consists of *t* vertices.

Having the user choose the number of vertices in $G_S$ and using a greedy method to solve the problem will almost certainly result in a reduced graph that with produce a suboptimal result when combined with an MR function. Thus, there is a need to develop a method that finds an underlying manifold of the graph and has a near-globally optimal solution.

## 4.3 Methodological Development

My semi-supervised method incorporates both feature selection and graph sampling to improve the accuracy and efficiency of model prediction by eliminating noisy or redundant instances and features. The base model that I use for semi-supervised regression is Laplacian Regularized least squares (LapRLS), a manifold regularization algorithm (Belkin et al. 2006). For feature selection I introduce a wrapper method called constrained particle swarm optimization-SSL (cPSO-SSL), and for graph sampling, I develop a sampling method called nearest neighbors graph reduction (NNGR).

### 4.3.1 Laplacian-Regularized Least Squares (LapRLS)

I adopted the graph-based SSL (Belkin et al. 2006) as the base learner model because of its proven high accuracy and efficiency in various semi-supervised applications, as well as its inductive learning ability that allows the trained model to be used to predict new patients. The formula for LapRLS is summarized as follows:

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \quad \frac{1}{L}\sum_{l=1}^{L}\left(y_l - f(\mathbf{x}_l)\right)^2 + \gamma_A \|f\|_K^2 + \frac{\gamma_I}{\sum_{i,j} w_{ij}} \mathbf{f}^T \mathbf{\Omega} \mathbf{f} \tag{4.2}$$

$L$ is the number of labeled instances in the training set. $y_l$ is the response value of the $l$-th instance. $\mathbf{x}_l$ are the predictive features for the $l$-th instance. $f(\mathrm{x}_l)$ is the predictive function of the $\mathbf{x}_l$. $(y_l - f(\mathbf{x}_l))^2$ is a loss function that measures the discrepancy between actual and predicted response. $f$ is a function on the reproducing kernel Hilbert space (RKHS), $\mathcal{H}_K$, with a Mercer kernel $K$. $\|f\|_K^2$ is a norm on $\mathcal{H}_K$, which encourages stability and generalizability of the solution. $\gamma_A$ is a tuning parameter. The graph encoded Laplacian matrix, $\mathbf{\Omega} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the vertex degree matrix, i.e., a diagonal matrix with

diagonal elements being the total sum of edge weights associated with each vertex, and **W** is the matrix of all the edge weights. The $\mathbf{f}^T\mathbf{\Omega f}$ term encourages instances that have similar **x**-values also have similar predictions. $\gamma_I$ is a tuning parameter and $\sum_{i,j} w_{ij}$ is used as a scaling factor. More discussion on this algorithm and its properties can be found in Belkin et al. 2006.

### 4.3.2 Constrained Particle Swarm Optimization (cPSO)

To perform feature selection, I utilize a modified version of a wrapper method called particle swarm optimization (PSO), originally developed as a population-based stochastic optimization technique, and then extended for feature selection in classification. I use constrained PSO (cPSO), which can honor a pre-specified maximum number of features to better avoid overfitting (Gaw et al. 2018). The advantage of using cPSO versus classical sequential forward selection is that it is more likely to find a near-global optimal solution.

In theory, cPSO can work with any classification or regression model. In this paper, I focus on LapRLS to demonstrate a first-of-its-kind semi-supervised feature selection method for regression problems. In cPSO, there are a number of particles in our solution space, for which each particle contains a potential feature set that is used for the solution. Each feature in each particle has a corresponding velocity

$$v_{id}^t = \omega^t v_{id}^{t-1} + c_1 r_1 (p_{id}^t - x_{id}^t) + c_2 r_2 \left(p_{gd}^t - x_{id}^t\right) \tag{4.3}$$

Where $v_{id}^t$ is the velocity of the $d^{\text{th}}$ dimension of the $i^{\text{th}}$ particle for $t^{\text{th}}$ iteration, $\omega^t$ is an inertia value, $c_1$ and $c_2$ are pre-defined constants, $r_1$ and $r_2$ are uniform(0,1) random

variables, $x_{id}^t$ is the current position of the particle, $p_{id}^t$ is the current best position for the individual particle, and $p_{gd}^t$ is the current best global position for the population of particles. The velocities are then put through a sigmoid function $S$, then they are ranked from highest to lowest. The top $k$ velocities are chosen as features in the particle as long as the feature has a good quality (i.e., $S(v_{id}^t) > 0.5$). The position formula is shown below.

$$x_{id}^{t+1} = \begin{cases} 1, & if \;\; d \leq k \; and \; S(v_{id}^t) > 0.5 \\ 0, & otherwise \end{cases} \tag{4.4}$$

More in-depth explanation for the mechanics of the cPSO algorithm can be found in Chapter 3 and Gaw et al. 2018.

### 4.3.3 Nearest Neighbors Graph Reduction (NNGR)

Let us suppose that there is a set of unlabeled and labeled instances that are connected to each other by a sparse graph, such that there is a density requirement that requires each instance to have at least $k$ nearest neighbors within a radius of constant length $\varepsilon$. To sample the graph such that the inherent manifold is retained (without having more points than necessary), the following heuristic can be used for each instance: either (1) keep the instance of interest, or (2) keep one of the instance's nearest neighbors within a fixed ε-radius.

This algorithm can be formulated by a simple integer programming (IP) problem that can be solved easily by common optimization heuristics, such as branch and bound (Land and Doig 1960).

$$\min_{\mathbf{b}} \sum_{i=1}^{u+l} b_i \tag{4.5}$$

$$s.t. \qquad \sum_{i=1}^{u+l} A_{ji} b_i > 1 \quad \forall\, j = 1, \dots, u+l \tag{4.5.1}$$

$$b_i = 1 \quad \forall\, i = 1, \dots, l \tag{4.5.2}$$

$$b_i \in \{0,1\} \quad \forall\, i = l+1, \dots, l+u \tag{4.5.3}$$

Where $\mathbf{b}_{(u+l)\times 1}$ is a vector that indicates whether each instance is included in the sample, and $A_{ji}$ is the $j^{\text{th}}$ row and $i^{\text{th}}$ column of **A**. **A** is a binary matrix that indicates connections between instances in the sparse graph (and includes self-connections for each node). Constraint (4.5.1) ensures that either instance $j$ ($j = 1, \dots, u+l$) or one of its nearest neighbors is included in the reduced graph. Constraint (4.5.2) forces all labeled instances to be included in the reduced graph (since labeled instances are few and therefore not disposable). Constraint (4.5.3) makes $b_i$ for the unlabeled instances constrained to 0 or 1 (i.e., whether or not the instance is included in the final graph).

A modification to (4.5.1) can give more flexibility to modulate the sparsity of the reduced graph. Instead of ensuring that only the instance of one of its nearest neighbors is included in the sample, there can be a heuristic that ensures that greater than $\lambda$ instances are sampled for each group of instance $j$ and its nearest neighbors ($j = 1, \dots, u+l$). This allows increased flexibility with sampling and provides another tuning metric that can potentially improve model accuracy.

The proposed Nearest Neighbors Graph Reduction (NNGR) method that incorporates the above modification is defined below:

$$\min_{\mathbf{b}} \sum_{i=1}^{u+l} b_i \tag{4.6}$$

$$s.t. \qquad \sum_{i=1}^{u+l} A_{ji} b_i > \lambda \quad \forall\, j = 1, \dots, u+l \tag{4.6.1}$$

$$b_i = 1 \quad \forall\, i = 1, \dots, l \tag{4.6.2}$$

$$b_i \in \{0,1\} \quad \forall\, i = l+1, \dots, l+u \tag{4.6.3}$$

(4.6), (4.6.2), and (4.6.3) are identical in formulation to (4.5), (4.5.2), and (4.5.3), respectively. Constraint (4.6.1) ensures that greater than $\lambda$ instances are sampled for each group of instance $j$ and its nearest neighbors ($j = 1, \dots, u+l$).

This method can be shown to be quite effective. Even with noisy data, there are simple pre-processing methods (e.g., covariance matrices, clustering techniques, etc.) that can eliminate most outliers (for example, Wang and Zhang 2008).

### 4.3.4 Standard Error Scree

Including too many features in the model can cause generalization issues with predicting new instances. By running cPSO at several different maximum feature settings (i.e., $k = 1, \dots, J$), one can generate a plot of model error at different maximum feature settings). If cPSO is forced to include the best solution from the previous $k$ for each maximum feature setting, the error will monotonically decrease as maximum features are tested. The standard error scree (SES) method was utilized to automatically determine the optimal number of features to choose for the model (Zoski and Jurs 1996). Originally used for scree plots, SES can be directly applied since scree plots also decrease monotonically.

Multiple linear regression analysis was utilized to find the optimal number of features for the model by solving for the standard error of estimate $s_{Y \cdot X}$ (described in the paragraph below) for each regression.

The regression line for each maximum number feature setting was determined (the maximum feature number as the predictor, $X$, and MAE as the target, $Y$). The results were tabulated as follows (1) the maximum feature settings used in the calculations (1 through $J$, 2 through $J$, …, $J-2$ through $J$), and (2) the standard error of each regression ($s_{Y \cdot X_1}$, $s_{Y \cdot X_2}$, …, $s_{Y \cdot X_{J-2}}$). Finishing the series of regressions with calculations for $J-2$ through $J$ is consistent with Cattell's first guideline (i.e., three sequential points form an undesirable low limit for drawing a scree plot). To calculate standard error, $s_{Y \cdot X} = \sqrt{\left(Y - \hat{Y}\right)^2/(K-2)}$ was used, where $Y$ is the error value, $\hat{Y}$ is the predicted value of regression, and $K$ is the largest maximum features setting that is in the test.

The value of $1/J$ was chosen as the threshold for the standard error from which to determine whether allowing more maximum number of features produces nontrivial improvement in the results. This threshold value is based on recommendation from Zoski and Jurs 1996. Thus, each SES corresponding to a maximum feature setting that exceeds $1/J$ indicates a nontrivial improvements, whereas values less than or equal to $1/J$ indicates trivial improvements.

4.3.5 Model framework

Figure 7 summarizes the main functionalities of the proposed s2SSL algorithm. The steps of the s2SSL algorithm are also summarized below:

***Step 1 (hyperparameter initialization)*:** Set the following hyperparameters for s2SSL,

- *s2SSL parameters*: $k = 1, \dots, J$: the range of maximum number of features to set for cPSO
- *cPSO parameters*: $I$, the number of particles; $T$, the maximum number of iterations for the particles
- *NNGR parameters*: $k_{min}$ and $k_{min}^{sample}$: the minimum number of instances for instances pre- and post-sampling to be connected in the graph; $\lambda$, the parameter that controls sampling sparsity
- *LapRLS parameters*: values of $\gamma_A$, $\gamma_I$, and $\eta$ for tuning in a graph search fashion

Iterate steps 2-4 with $k = 1 \dots J$.

***Step 2 (cPSO initialization)*:** Set the initial position of the $i$-th particle, $\mathbf{x}_i^0$, by randomly choosing $k$ elements in $\mathbf{x}_i^0$ to be one while making other elements to be zero. If $k > 1$, include the global best solution found in $k - 1$ as one of the best particles. This ensures that cPSO at the current $k$ obtains a solution that is at least as good as the previous $k - 1$. Use the features corresponding to the non-zero elements in $\mathbf{x}_i^0$ to compute an error on the validation set, $f(\mathbf{x}_i^0)$—error is calculated by the ***NNGR + LapRLS sub-step*** (defined below). Set the initial velocity, $\mathbf{v}_i^0$, by sampling each element $\mathbf{v}_i^0$ from $U[-V_{max}, V_{max}]$. Use (4.4) to update the initial position of each particle and get $\mathbf{x}_i^1$. Iterate Steps 3-4 with $t = 1, \dots, T$.

***Step 3 (velocity updating)*:** Examine all previous positions of the $i$-th particle, $f(\mathbf{x}_i^0),\ldots, f(\mathbf{x}_i^{t-1})$ and find the position giving the smallest validation error, $\mathbf{p}_i^t$. Examine the current positions of all the particles, $f(\mathbf{x}_i^t),\ldots, f(\mathbf{x}_I^t)$, and find the position giving the smallest validation error, $\mathbf{p}_g^t$. Sample $r_1$ and $r_2$ from $U[0,1]$. Use (4.3) to compute the velocity $\mathbf{v}_i^t$. If $v_{id}^t > V_{max}$, set $v_{id}^t = V_{max}$; if $v_{id}^t < -V_{max}$, set $v_{id}^t = -V_{max}$.

***Step 4 (position updating)*:** Order the elements in $\mathbf{v}_i^t$ from the largest to the smallest. Use (4.4) to compute the new position $\mathbf{x}_i^{t+1}$. If the maximum number of iterations has been reached—i.e., $t+1 = T$—examine the current positions of the particles, $f(\mathbf{x}_i^{t+1}),\ldots, f(\mathbf{x}_I^{t+1})$, and output the position giving the smallest validation error as the optimal solution, together with the corresponding validation error and the features that are selected. Otherwise, go back to step 3.

***Step 5 (determining optimal maximum features setting)*:** Using the standard error scree method, the optimal maximum feature setting is found by applying a regression line to the error of each maximum feature setting and selecting the $k$ such that the standard error of the regression is less than $1/J$. Output the corresponding position as the optimal solution, together with the corresponding validation error and the features that are selected.

***NNGR + LapRLS sub-step*:** Generate a graph that ensures each instance has at least $k_{min}$ nearest neighbors, and then apply NNGR sampling. With the sampled subset, create a graph Laplacian that ensures each instance has at least $k_{min}^{sample}$ nearest neighbors.

Apply LapRLS to the samples and optimize the error on a separate validation set by performing a graph search with varying values of $\gamma_A, \gamma_I$, and $\eta$.

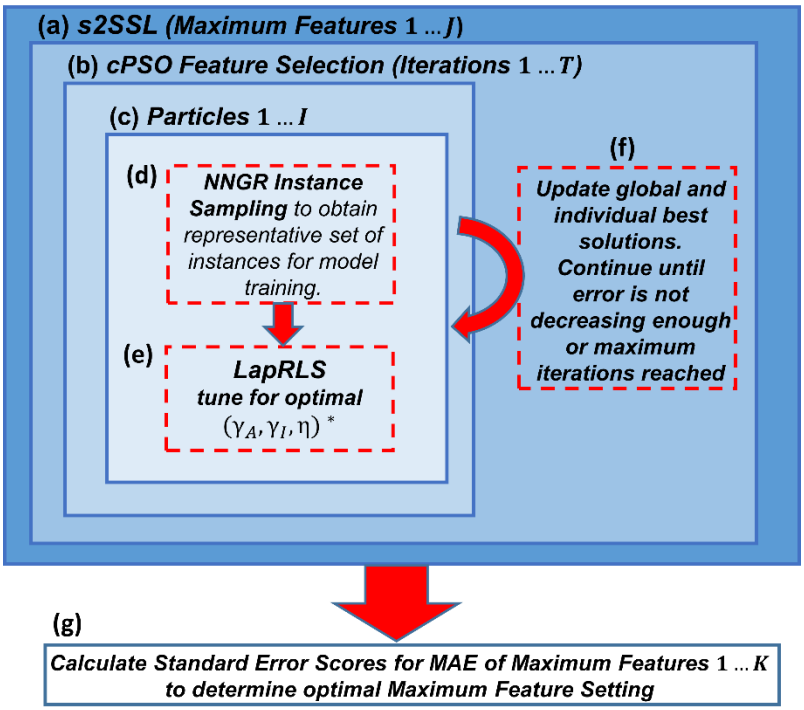



Figure 7: The framework of s2SSL. (a) s2SSL is ran at several different settings of maximum features (1 ... $J$ maximum features). (b) cPSO is initialized and set to run for $T$ maximum iterations. (c) For each of the $p$ particles in the cPSO algorithm there is a potential feature subset solution that is tested. (d) In each particle, NNGR instance sampling selects a sample of instances for model training. (e) LapRLS is then trained on this sample to minimize the error by tuning $\gamma_A$, $\gamma_I$, and $\eta$. (f) The global and individual best solutions are then updated, and cPSO continues until error is no longer decreasing enough or maximum iterations are reached. (g) Finally, the standard error scores are calculated for MAE of maximum features 1 ... $K$ to determine the optimal maximum feature setting.

## 4.4 Simulation Tests

### 4.4.1 Simulation: Overview

Simulation data was generated according to the make_s_curve function found on sci-kit learn

(https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_s_curve.html).

Appendix B at the end of this work provides details how the s curve was calculated. For comparing the utility of the semi-supervised learning baseline model (LapRLS) to its supervised learning counterpart (RLS), I generated an s curve with s-curve noise = 0.15 and 150 data instances shown in Figure 8, where there were 6 labeled instances (in color) and 144 unlabeled instances (in gray). The range of labeled instances was from -5 (blue) to +5 (yellow).

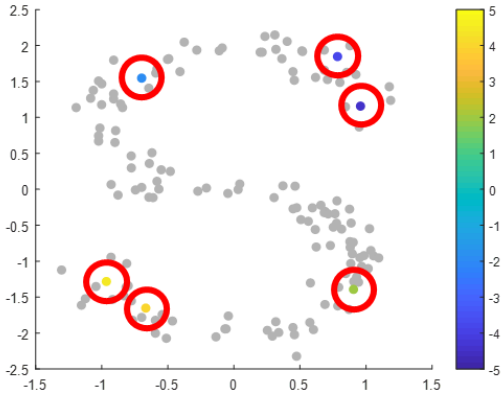



Figure 8: S curve used in model training to compare utility of semi-supervised over supervised learning. There are 144 unlabeled instances (in gray) and 6 labeled instances (in color and circled in red). The range of the labeled instances is from -5 (blue) to +5 (yellow).

The error metric used for simulation tests is the mean absolute error (MAE), which is defined as MAE $= \sum_{i=1}^{n} |\hat{y}_i - y_i|$. $\hat{y}_i$ is the predicted value of instance $i$, $y_i$ is the true value of instance $i$, and $n$ is the total number of instances of in the set of response values $\mathbf{y}_{n\times 1}$. For model training, hyperparameters were tuned to minimize the MAPE of an independently generated validation set consisting of 25 instances. The tuning parameters and ranges are $\gamma_A = [1 \times 10^{-3}, \ldots, 1 \times 10^{1}], \gamma_I = [1 \times 10^{-3}, \ldots, 1 \times 10^{1}], \eta =$ [1,2.5,4,5.5,7,8.5,10], $k_{min} = 6$, $k_{min}^{sample} = 4$.

4.4.2 Simulation: Semi-supervised Learning vs. Supervised Learning

Below in Table 8 is a comparison of the performance of s2SSL using LapRLS as the base classifier versus RLS. In the validation results, LapRLS performed significantly better than RLS both in terms of MAE and Pearson Correlation ($p = 0.00271$ and $p = 6.35 \times 10^{-4}$). Figure 9 shows the graph generated for the LapRLS model.

Table 8: Comparison of the LapRLS and RLS base models. Mean absolute error (MAE ± standard deviation) and Pearson correlation were used to compare the methods.

| | LapRLS | | RLS | |
|---|---|---|---|---|
| | **Training** | **Validation** | **Training** | **Validation** |
| **MAE** | 0.453 ± 0.310 | 0.293 ± 0.249 | 0.661 ± 0.585 | 0.659 ± 0.578 |
| **Correlation** | 0.983 | 0.993 | 0.941 | 0.949 |

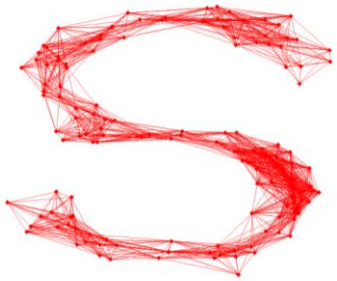


Figure 9: Graph generated by LapRLS such that each instance has at least $k_{min} = 6$ nearest neighbors

4.4.3 Simulation: Feature Selection

Next, the utility of the graph sampling algorithm in distinguishing between trivial and nontrivial features was examined. Tests were ran adding different numbers of noise (trivial) features such that each noise feature was distributed according to $N(0,5)$. For this section 1, 5, 10, 50, and 100 noise features were added to the dataset of two nontrivial

features to observe (1) the effect of adding noise on s2SSL without feature selection, and (2) the subsequent utility of cPSO feature selection.

cPSO was ran at several different maximum feature settings (from $k = 1, \ldots, 5$ maximum features), and the minimum error was chosen. In more sophisticated tasks, where greater than 2 nontrivial features are expected it is recommended to use a more sophisticated method to select features such as SES (Zoski and Jurs 1996), which will be used to select the optimal maximum feature setting in the application section. Table 9 summarizes the results from different feature settings. As more noise features were added, the performance of the model deteriorated. However, s2SSL with cPSO feature selection was able to successfully select the nontrivial features (summarized in the first row of Table 9). This is partially due to the deterioration of the graph when adding more noise features. Figure 10 demonstrates the graph deterioration as more features were added. In the validation results, cPSO performed better than no feature selection in terms of MAE and Pearson Correlation across all noisy feature settings ($p < 5 \times 10^{-5}$ in all cases).

Table 9: Comparison of LapRLS with cPSO feature selection to LapRLS without feature selection. Accuracy metrics are mean absolute error (MAE ± standard deviation) and Pearson Correlation. In all noise feature settings, cPSO chose the correct nontrivial features (results summarized in the first row). The remaining table shows how the performance without feature selection deteriorates as more noise features are added.

| | | Mean Absolute Error (MAE) | | Pearson Correlation | |
|---|---|---|---|---|---|
| | Noise Features | Training | Validation | Training | Validation |
| Feature Selection | All Settings | 0.453 ± 0.310 | 0.293 ± 0.249 | 0.983 | 0.993 |
| No Feature Selection | 1 | 0.821 ± 0.549 | 0.875 ± 0.573 | 0.925 | 0.923 |
| | 5 | 1.251 ± 0.952 | 1.481 ± 0.825 | 0.829 | 0.804 |
| | 10 | 1.655 ± 1.127 | 1.863 ± 1.081 | 0.637 | 0.608 |
| | 50 | 2.029 ± 1.355 | 1.805 ± 1.224 | 0.384 | 0.651 |
| | 100 | 2.117 ± 1.316 | 2.200 ± 1.317 | 0.301 | 0.305 |

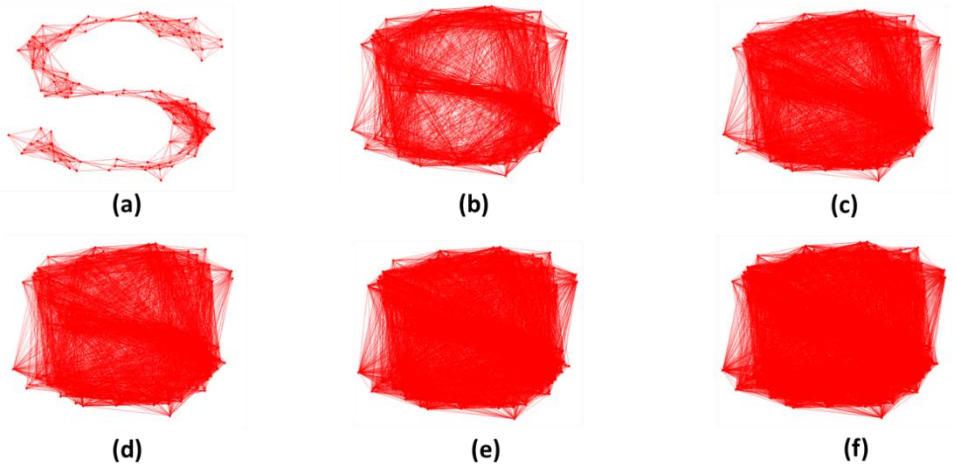



Figure 10: The deterioration of the graph between different instances as more noise features were added. The results present are for (a) no added noise features, (b) 1 added noise feature, (c) 5 added noise features, (d) 10 added noise features, (e) 50 added noise features, and (f) 100 added noise features. The deterioration of the graph reduces model performance.

4.4.4 Simulation: Graph-based sampling

In the next set of tests, I generated an S curve using s-curve noise = 0.15 and 2000 data instances. Figure 11 below shows the S curve that was generated, where there 1994 unlabeled instances (in gray) and were 6 labeled instances (in color). The range of labeled instances was from -5 (blue) to +5 (yellow).

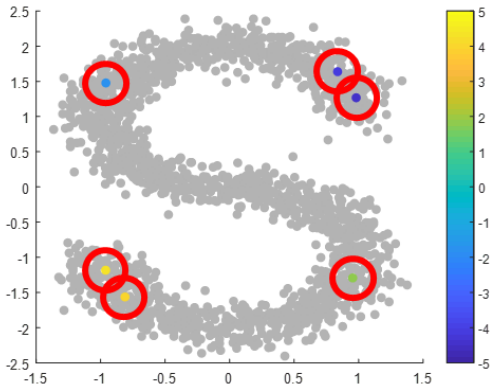



Figure 11: S curve used in model training to compare utility of sampling over not sampling. There are 1994 unlabeled instances (in gray) and 6 labeled instances (in color and circled in red). The color bar for the labeled date ranges from -5 (blue) to +5 (yellow).

Because there is a large number of instances in this dataset, to train my model in an efficient matter, it is necessary to use a sampling technique to reduce the number of instances to train. To achieve this purpose, the NNGR graph sampling technique (proposed in the Methodology section in 4.3.3) was employed.

Table 10 compares the result of LapRLS + NNGR at different settings of $\lambda$ to better understand $\lambda$'s effect on sampling sparsity and accuracy. The result for without sampling is also included as the baseline example. As one can observe, sampling greatly improves the (sample +) train + test times time (for LapRLS + NNGR, the time to sample is also included for a fair comparison). The best sampling result was found to be at $\lambda = 1$. For this setting, the validation MAE and Pearson Correlation were better than LapRLS without

sampling. When compared, Pearson Correlation was not found to be significantly different ($p = 0.437$), however improvement in MAE approached significance ($p = 0.0818$). This indicates that sampling may not only have capabilities to improve model train + test time, but also the model performance.

Table 10: Comparison of LapRLS + NNGR at different levels of $\lambda$. Training and validation errors are in terms of mean absolute error (MAE ± standard deviation) and Pearson Correlation, time refers to the time to sample + train + test, number sampled is the number of instances sampled from the dataset. The best sampling result is $\lambda = 1$ (in bold).

| | | Mean Absolute Error (MAE) | | Pearson Correlation | | | |
|---|---|---|---|---|---|---|---|
| | $\lambda$ | Train | Validation | Train | Validation | Time (s) | Number Sampled |
| **No Sampling** | **N/A** | 0.480 ± 0.340 | 0.495 ± 0.343 | 0.984 | 0.984 | 41.9 | N/A |
| **Sampling** | **0** | 0.524 ± 0.390 | 0.506 ± 0.375 | 0.982 | 0.986 | 8.5 | 65 |
| | **1** | **0.439 ± 0.332** | **0.363 ± 0.315** | **0.985** | **0.990** | **12.4** | **106** |
| | **2** | 0.465 ± 0.343 | 0.460 ± 0.331 | 0.985 | 0.987 | 6.6 | 169 |
| | **3** | 0.603 ± 0.490 | 0.568 ± 0.516 | 0.971 | 0.972 | 8.9 | 235 |
| | **4** | 0.557 ± 0.400 | 0.594 ± 0.438 | 0.978 | 0.976 | 8.9 | 236 |
| | **5** | 0.591 ± 0.437 | 0.611 ± 0.458 | 0.974 | 0.974 | 9.7 | 285 |
| | **6** | 0.599 ± 0.451 | 0.611 ± 0.467 | 0.973 | 0.972 | 7.7 | 309 |

Figure 12 shows the original graph made on the 2000 instances before graph sampling. As one can observe, the graph is very dense. This graph density is the major cause of the heavy computational cost of training s2SSL without sampling.

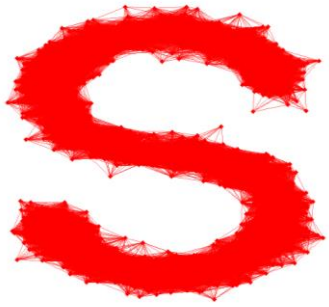


Figure 12: Graph made on 2000 instances (with $k_{min} = 6$) before sampling.

Figure 13 summarizes the different graphs made on the instances post sampling at different levels of $\lambda$ (all with $k_{min}^{sample} = 4$). One can observe the gradual change in the shape of the graph (moderately lower values of lambda, especially at $\lambda = 1$ to capture the underlying shape of the 'S' curve better than the higher values of lambda, which have some discontinuities in the graph due to the greater number of sampled instances when $\lambda = 5$ and $\lambda = 6$).

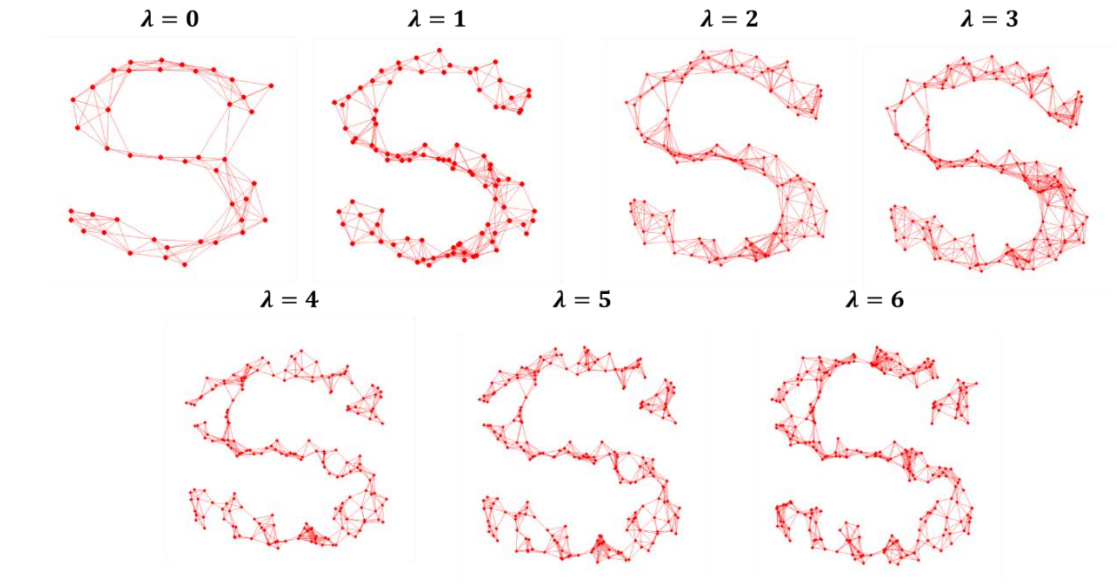



Figure 13: Graphs made on sampled instances post sampling at different levels of $\lambda$ (all with $k_{min}^{sample} = 4$).

4.4.5 Simulation: Feature selection + Graph-based sampling

Additional tests were performed to show feature selection and graph-based sampling working together to demonstrate the ability of s2SSL to build a model by selecting the most relevant features and instances. For all tests in this section, $\lambda = 1$ for graph-based sampling, since this value produced the best result in the previous section. Table 11 below demonstrates the performance of s2SSL at various numbers of added noise features with the model trained on the 2000 instance S curve dataset. The performance of the base model, LapRLS (without sampling or feature selection) is compared in all settings. s2SSL performed better than LapRLS in terms of validation MAE and Pearson Correlation for all noise feature settings ($p < 0.001$ for all cases).

Table 11: Summary of the performance of s2SSL versus the LapRLS baseline (without feature or sample selection) at several different added noise feature settings. Mean Absolute Error (MAE) ± standard deviation and Pearson Correlation are the accuracy metrics. s2SSL successfully chose the correct features at all noise feature settings and results for all tests are summarized in the first result row.

| | | Mean Absolute Error (MAE) | | Pearson Correlation | |
|---|---|---|---|---|---|
| | **Noise Features** | **Training** | **Validation** | **Training** | **Validation** |
| **s2SSL** | **All Settings** | 0.439 ± 0.332 | 0.363 ± 0.315 | 0.985 | 0.990 |
| **LapRLS** | **1** | 1.016 ± 0.832 | 1.088 ± 1.025 | 0.895 | 0.877 |
| | **5** | 1.269 ± 0.962 | 1.129 ± 0.905 | 0.824 | 0.847 |
| | **10** | 1.537 ± 1.193 | 1.987 ± 1.270 | 0.729 | 0.661 |
| | **50** | 2.286 ± 1.329 | 2.240 ± 1.411 | 0.345 | 0.311 |
| | **100** | 2.378 ± 1.364 | 2.276 ± 1.416 | 0.249 | 0.552 |

Table 12 summarizes the (sample +) train + test times for s2SSL versus LapRLS (sampling time was included in the calculation for s2SSL to allow for a fair comparison to LapRLS). As one can see, at relatively low number of features, the times are not very

different from each other. However, as one increases the number of noise features, the sample + train + test time for s2SSL explodes. In the age of parallel computing and high computer processing power, this does not become a large issue since processes can run in parallel on a computing cloud.

Table 12: Summary of (sample +) train + test times for s2SSL versus LapRLS compared at different numbers of noise features.

| Noise Features | s2SSL Sample + Train + Test Time (s) | LapRLS Train+ Test Time (s) |
|---|---|---|
| **1** | 128.1 s | 74.0 s |
| **5** | 127.0 s | 92.5 s |
| **10** | 109.9 s | 98.7 s |
| **50** | 197.8 s | 80.5 s |
| **100** | 318.7 s | 85.7 s |

## 4.5 Application to Parkinson's Disease (PD) Telemonitoring

### 4.5.1 Parkinson's Disease Telemonitoring: Background

In this section, I demonstrate the utility of s2SSL to building parsimonious models to predict disease severity of PD patients using features collected from the mPower app installed on patient's iPhones. Utilizing the Apple ResearchKit library, Sage Bionetworks released the mPower app in March 2015 for an observational study on smartphone-based telemonitoring of PD (Bot et al. 2016). The mPower app obtains information from daily exercises performed by patients with the purpose of monitoring PD disease progression.

To participate in the mPower study, each participant had to self-navigate through eligibility criteria (i.e., age at least 18 years, U.S. Resident, comfortability with reading and writing English on the iPhone) and submit e-consent to the conditions. The study was performed in accordance with the Western Institutional Review Board. Once the consent

process is finished, users were presented with the option of performing four different activities in the app—namely, 'tapping', 'voice', 'memory', and 'walking'—each of which can be performed at most three times per day. Among the available studies, I extract features from the tapping and voice studies since both capture well-known symptoms of PD (tapping: Lainscek et al. 2012, Lee et al. 2016; voice: Holmes et al. 2000, Skodda et al. 2009, Chattopadhyay et al. 2012).

The purpose of the tapping study is to measure speed and dexterity of each user's tapping ability. Users are instructed to use two fingers on the same hand to tap alternately between two fixed points on the screen for a period of 20 seconds. To generate features, time series signals are collected by the accelerometer and touch screen on the smartphone.

The voice study made recordings of user's sustained phonation by instructing users to say 'Aaaaah' into the microphone at a steady volume for at most 10 s. Included in the data for this activity are the audio files that contain measures from the iPhone microphone for 10 s of phonation. Using the Voice Analysis Toolbox (available at: https://people.maths.ox.ac.uk/tsanas/software.html), features were processed for the objective characterization of the user's voice (Tsanas et al. 2010). The features generated from the Voice Analysis Toolbox are mainly directed at quantifying amplitude (shimmer variants), frequency (jitter variants) and increased noise (signal-to-noise measures).

The most popular metric used to quantify the severity of PD is UPDRS, which is a summary score from a survey administered to patients. Recently, the Movement Disorder Society UPDRS (MDS-UPDRS) was developed by the Movement Disorder Society to address a number of ambiguities, weaknesses, and areas of inclusions needed in light of new scientific developments (Goetz et al. 2008). The MDS-UPDRS is a summary score

from a subset of the questions used in the UPDRS survey. An MDS-UPDRS score of 0 denotes no disability, while an MDS-UPDRS score of 64 indicates the worst possible disability. Usually, the MDS-UPDRS score is obtained in a specialized clinic, which requires the patient to be physically present during testing. It will be shown that there is capability to accurately predict the MDS-UPDRS score using smartphone-collected tapping and voice signals. To enable this proof of concept, actual MDS-UPDRS scores from the clinic were collected from each user on a monthly basis. Typically, daily MDS-UPDRS scores are obtained through linear interpolation, as a linear trend of UPDRS has been validated in previous works (Chan and Holford 2001, Schüpbach et al. 2009, Tsanas et al. 2010). However, it would be better to train models based on the monthly collected ground truth MDS-UPDRS scores instead of relying on approximated values. Thus s2SSL will be utilized to train models only needing a few labeled instances and select features from the set of those that are available.

### 4.5.2 Parkinson's Disease Telemonitoring: Dataset Description

A subset of 37 PD patients were included in the current study. These patients were selected on the basis of having monthly MDS-UPDRS scores for at least three months as well as complete daily tapping and voice information.

43 tapping features were extracted from the tapping time series data, based on previous studies (Taylor et al. 2005, Arora et al. 2015, Kassavetis et al. 2016). Taylor et al. connects the UPDRS motor score with kinematics of an alternating finger-tapping task using features generated from Quantitative digitography (QDG) (Taylor et al. 2005). Arora et al. presents a summary of measures to quantify tremor, fatigue, tapping speed, inter-tap

interval and tapping speed from using time series finger tapping data (Arora et al. 2015). Kassavetis et al. also presents several tapping-related features (Kassavetis et al. 2016).

339 voice features were extracted from the voice time series data, based on previous studies (Tsanas et al. 2010, Yoon and Li 2019). Tsanas et al. proposed a number of novel signal processing algorithms for speech signals collected from at-home-testing devices (AHTDs) (Tsanas et al. 2010). They utilized robust feature selection algorithms to select the voice measures as input to non-parametric regression and classification algorithms to predict the UPDRS score. Yoon and Li built a positive transfer learning model to develop patient-wise predictions on voice features generated from AHTDs (Yoon and Li 2019).

s2SSL was trained on three different datasets: (1) tapping, (2) voice, (3) tapping + voice combined. The reason why I decided to test on combined datasets of tapping and voice is because there is significant variability in the presentation and progression of PD symptoms (Bot et al. 2016) across patients, and I hypothesize that having a model trained on different types of PD symptoms will result in significantly improved results. Tuning on the labeled data in the validation set was used to optimize the parameters $\gamma_A = [1e \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}]$, $\gamma_I = [1 \times 10^{-3}, 5 \times 10^{-3}, \ldots, 1 \times 10^{1}, 5 \times 10^{1}]$ and $\eta = [1,2.5,4,5.5,7,8.5,10]$.

To expedite the training process, computing was performed using two Intel Xeon E5-2680 v4 CPUs running at 2.40 GHz, which provide 28 CPU cores to perform calculations for each particle in cPSO in parallel. The advantage of using an evolutionary algorithm, such as cPSO, is that utilizing parallel computing is straightforward to implement and easy to scale for datasets with a large number of features.

For this data set, labeled samples are those that have both features collected from mobile data and MDS-UPDRS scores collected on the same day. Unlabeled samples have features collected from mobile data but no MDS-UPDRS scores. For each dataset, two labeled samples were randomly selected from each patient such that they were selected from different days. Ensuring that samples were collected from different days avoid a potential breakdown in the "similar $X \rightarrow$ similar $Y$" assumption that is required by the semi-supervised base model, LapRLS (i.e., in order for LapRLS to work properly, instances with similar $X$-values must have similar $Y$-values; for additional information regarding this property, refer to Belkin et al. 2006).

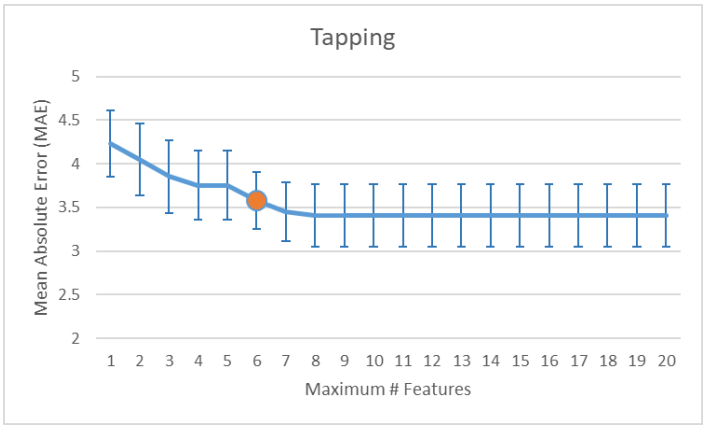
To test the generalizability of the model, a validation set was made from the remaining labeled samples such that all labeled samples were selected from different days. Unlabeled samples were selected randomly once per week. For all tests, the training set contained a total of 563 unlabeled instances and 74 labeled instances, while the validation set contained 70 instances.

### 4.5.3 Parkinson's Disease Telemonitoring: Accuracy results on semi-supervised regression feature selection task

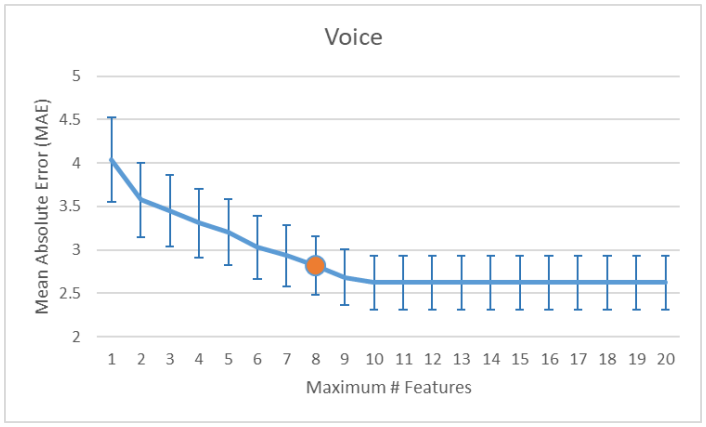
s2SSL was tested on several different maximum feature settings of cPSO (from 1-20 maximum features allowed to be selected for model training). The standard error scree (SES) method, described in section 4.3.4, was used to determine the optimal feature number. Figure 14 and Table 13 show the elbow plots of MAE along with their

corresponding tables of MAE's and SES's. Tapping, voice, and tapping + voice chose 6, 8, and 10 features to include in the final model, respectively.

**(a)**




**(b)**




**(c)**

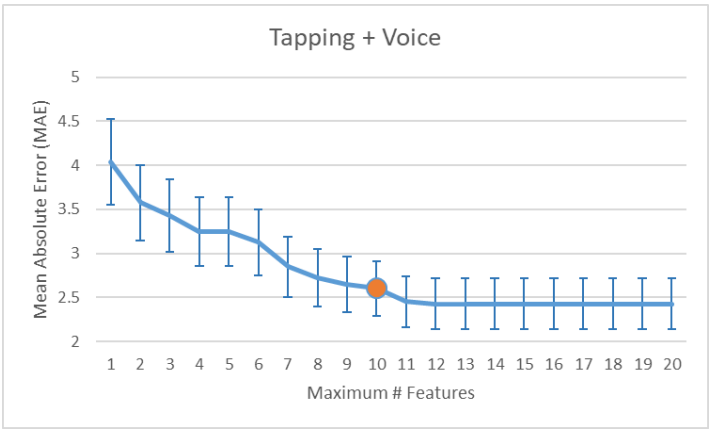



Figure 14: The elbow plots of the mean absolute error (MAE ± standard error) for (a) tapping features, (b) voice features, and (c) tapping + voice features. The orange dot indicates the place for which the optimal maximum features setting was selected. Tapping, voice, and tapping + voice chose 6, 8, and 10 features to include in the final model, respectively.

Table 13: The resulting mean absolute error (MAE ± standard error) and standard error scree (SES) for several different maximum feature settings of s2SSL applied to (a) tapping features, (b) voice features, and (c) tapping + voice features. The chosen maximum feature number threshold is in bold and marked with a '*'. This chosen maximum feature number threshold corresponds with the first SES that is less than 1/20.

**(a) Tapping Features**

| Maximum # Features | MAPE | SES |
|---|---|---|
| **1** | 4.235 ± 0.379 | 0.159 |
| **2** | 4.049 ± 0.410 | 0.129 |
| **3** | 3.856 ± 0.417 | 0.103 |
| **4** | 3.751 ± 0.397 | 0.0891 |
| **5** | 3.751 ± 0.397 | 0.0777 |
| **6** | **3.580 ± 0.328*** | **0.0402*** |
| **7** | 3.452 ± 0.338 | 0.011 |
| **8** | 3.407 ± 0.355 | 3.28E-16 |
| **9** | 3.407 ± 0.355 | 1.28E-15 |
| **10** | 3.407 ± 0.355 | 9.24E-16 |
| **11** | 3.407 ± 0.355 | 5.87E-16 |
| **12** | 3.407 ± 0.355 | 3.75E-16 |
| **13** | 3.407 ± 0.355 | 5.13E-16 |
| **14** | 3.407 ± 0.355 | 0 |
| **15** | 3.407 ± 0.355 | 9.42E-16 |
| **16** | 3.407 ± 0.355 | 1.40E-15 |
| **17** | 3.407 ± 0.355 | 4.44E-16 |
| **18** | 3.407 ± 0.355 | 0 |
| **19** | 3.407 ± 0.355 | N/A |
| **20** | 3.407 ± 0.355 | N/A |

**(b) Voice Features**

| Maximum # Features | MAPE | SES |
|---|---|---|
| **1** | 4.040 ± 0.483 | 0.230 |
| **2** | 3.577 ± 0.428 | 0.177 |
| **3** | 3.453 ± 0.413 | 0.161 |
| **4** | 3.308 ± 0.395 | 0.141 |
| **5** | 3.202 ± 0.383 | 0.122 |
| **6** | 3.030 ± 0.362 | 0.0941 |
| **7** | 2.932 ± 0.350 | 0.0742 |
| **8** | **2.823 ± 0.337*** | **0.0487*** |
| **9** | 2.683 ± 0.321 | 0.0161 |
| **10** | 2.623 ± 0.314 | 0.000102 |
| **11** | 2.623 ± 0.313 | 3.14E-16 |
| **12** | 2.623 ± 0.313 | 2.37E-16 |
| **13** | 2.623 ± 0.313 | 4.80E-16 |
| **14** | 2.623 ± 0.313 | 5.25E-16 |
| **15** | 2.623 ± 0.313 | 2.22E-16 |
| **16** | 2.623 ± 0.313 | 4.44E-16 |
| **17** | 2.623 ± 0.313 | 5.44E-16 |
| **18** | 2.623 ± 0.313 | 7.69E-16 |
| **19** | 2.623 ± 0.313 | N/A |
| **20** | 2.623 ± 0.313 | N/A |

Table 13 (continued)

**(c) Tapping + Voice Features**

| Maximum # Features | MAPE | SES |
|---|---|---|
| **1** | 4.040 ± 0.483 | 0.234 |
| **2** | 3.577 ± 0.428 | 0.188 |
| **3** | 3.428 ± 0.410 | 0.176 |
| **4** | 3.245 ± 0.388 | 0.164 |
| **5** | 3.245 ± 0.388 | 0.158 |
| **6** | 3.127 ± 0.374 | 0.132 |
| **7** | 2.847 ± 0.340 | 0.0881 |
| **8** | 2.725 ± 0.326 | 0.0712 |
| **9** | 2.648 ± 0.316 | 0.0591 |
| **10** | **2.600 ± 0.311*** | **0.0457*** |
| **11** | 2.450 ± 0.293 | 0.00711 |
| **12** | 2.425 ± 0.290 | 0 |
| **13** | 2.425 ± 0.290 | 6.01E-16 |
| **14** | 2.425 ± 0.290 | 7.16E-16 |
| **15** | 2.425 ± 0.290 | 3.14E-16 |
| **16** | 2.425 ± 0.290 | 2.56E-16 |
| **17** | 2.425 ± 0.290 | 8.31E-16 |
| **18** | 2.425 ± 0.290 | 4.44E-16 |
| **19** | 2.425 ± 0.290 | N/A |
| **20** | 2.425 ± 0.290 | N/A |

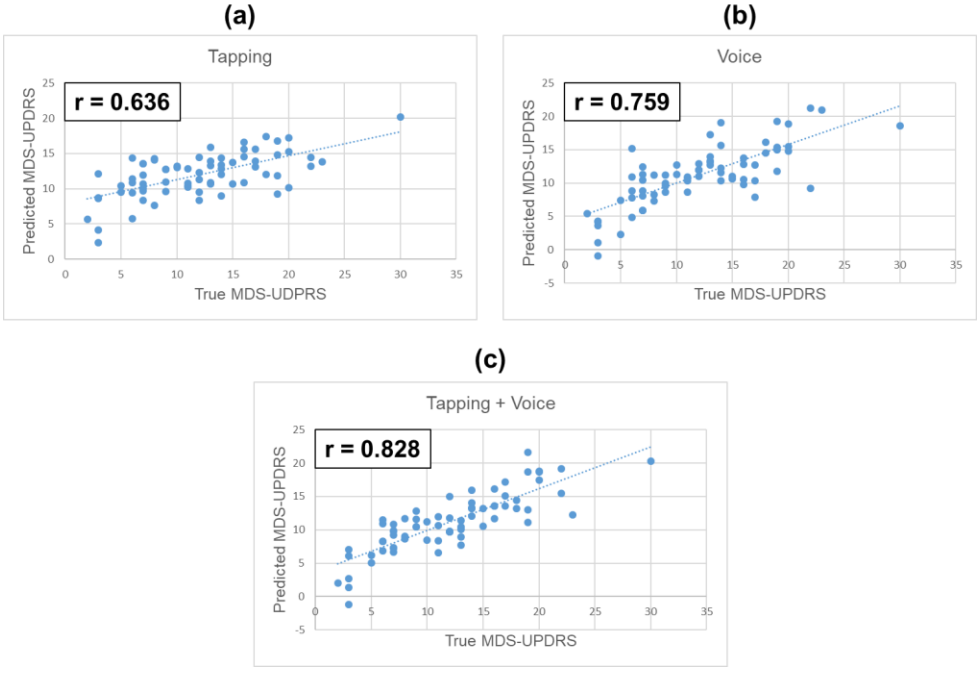
Table 14 provides a summary of the performance of the final models chosen by SES for s2SSL trained on tapping, voice, and tapping + voice features. s2SSL trained on tapping and voice features combined was not found to be significantly better than s2SSL trained on voice features only in both MAE and Pearson Correlation ($p = 0.29$ and $p = 0.27$). s2SSL trained on voice features performs significantly better than s2SSL trained on tapping features in terms of MAE ($p < 0.05$), but not in terms of Pearson correlation ($p = 0.16$). However, when s2SSL is trained on tapping and voice features combined, performance is significantly improved both in terms of MAPE and Pearson Correlation ($p = 0.01$ for both), indicating that including features from both modalities (tapping and voice) provides a greater improvement on the results than using either feature modality

alone. Figure 15 shows the scatter plots of predicted MDS-UPDRS versus true MDS-UPDRS for s2SSL trained on tapping, voice, and tapping + voice features.

Table 14: Summary of the final models chosen by s2SSL for tapping features, voice features, and tapping + voice features. MAE (training) and MAE (validation) refer to the mean absolute errors on the training and validation sets, respectively; Correlation (training) and Correlation (validation) refer to the Pearson correlation on the training and validation sets, respectively; and sample + train + test time is the total time required to sample, train, and validate the models.

| | MAE (training) | Correlation (training) | MAE (validation) | Correlation (validation) | Sample + Train + Test time (s) |
|---|---|---|---|---|---|
| **Tapping** | 3.448 ± 0.346 | 0.717 | 3.580 ± 0.328 | 0.636 | 590.1 |
| **Voice** | 2.119 ± 0.181 | 0.936 | 2.823 ± 0.337 | 0.759 | 644.8 |
| **Tapping + Voice** | 1.850 ± 0.192 | 0.933 | 2.600 ± 0.311 | 0.828 | 598.8 |

Figure 15: Scatter plots of predicted MDS-UPDRS score versus true MDS-UPDRS score for (a) tapping, (b) voice, and (c) tapping + voice features. Tapping + voice produces the best correlation results.



### 4.5.3 Parkinson's Disease Telemonitoring: Biomarker Identification

Table 15 summarizes the features chosen by s2SSL when trained on tapping, voice, and tapping + voice features. Table 16 provides definitions for each feature chosen. Additional information about the meaning of the chosen tapping and voice features can be found in and https://github.com/Sage-Bionetworks/mpowertools/blob/master/FeatureDefinitions.md and (Arora et al. 2015), respectively. Since s2SSL trained on tapping + voice features achieved the highest accuracy, I examine this result more closely.

The tapping features chosen were the *kurTapInter* and *madDriftRight*. *kurTapInter* is the kurtosis of the inter-tap interval. *madDriftRight* is the median absolute deviation of drift of finger position between consecutive taps in the right button. PD patients have been found to have a higher intra-individual variability of finger tapping due to a lack of control in fine motor capabilities (Roalf et al. 2018).

The voice features chosen fit under three categories:

**(1) Shimmer** (*Shimmer->F0_PQ5_classical_Schoentgen* and *Shimmer->F0_FM*),

**(2) Mel Frequency Cepstral Coefficients (MFCCs)** (*mean_MFCC_1st*, *mean_MFCC_6th*, and *std_8th_delta_delta*), and

**(3) Wavelet measures** (*det_TKEO_mean_10_coef*, *app_LT_entropy_log_1_coef*, and *app_LT_entropy_log_5_coef*).

The shimmer (cycle-to-cycle variation in amplitude) of voice signal is known to be higher in PD patients than healthy controls (Ramig et al. 1988, Hertrich et al. 1995, Jiang et al. 1999). Shimmer has frequently been used as a measure of voice signal for PD. Mel Frequency Cepstral Coefficients (MFCCs) capture variation in both vocal folds and the

vocal tract (i.e., tongue, lips, jaw, etc.). PD research has demonstrated that, in addition to the vocal folds that traditional measures capture, articulators of the vocal tract (i.e., tongue, lips, jaw, etc.) are affected by the disease (Ho et al. 1998). Wavelet measures are derived from the discrete wavelet transform (DWT), which can quantify both regularity effects (scale aspects) and transient processes (time aspects) (Tsanas 2012). DWT decomposes the wavelet signal into detail information (detail coefficients) and course approximation (approximation coefficients). The main rationale for wavelet measures is that people with pathological voices cannot sustain a vowel with minimum deviation from exact periodicity, while healthy controls can (Titze 2000).

Table 15: Features chosen for s2SSL training on (a) tapping features, (b) voice features, and (c) tapping + voice features.

| (a) Tapping | (b) Voice | (c) Tapping + Voice |
|---|---|---|
| iqrTapInter<br>ar2TapInter<br>meanDriftLeft<br>kurDriftLeft<br>skewDriftRight<br>madDriftRight | Shimmer->F0_dif_percent<br>Shimmer->F0_PQ11_classical_Baken<br>Shimmer->F0_abs0th_perturb<br>VFER->SNR_TKEO<br>std_MFCC_6th<br>det_entropy_log_4_coef<br>app_entropy_shannon_6_coef<br>det_LT_entropy_log_4_coef | **Tapping:**<br>kurTapInter<br>madDriftRight<br>**Voice:**<br>Shimmer->F0_PQ5_classical_Schoentgen<br>Shimmer->F0_FM<br>mean_MFCC_1st<br>mean_MFCC_6th<br>std_8th_delta_delta<br>det_TKEO_mean_10_coef<br>app_LT_entropy_log_1_coef<br>app_LT_entropy_log_5_coef |

Table 16: Description of features chosen for (a) Tapping, (b) Voice, and (c) Tapping + Voice.

**(a) Tapping**

| Feature | Definition |
|---|---|
| iqrTapInter | Interquartile Range of inter-tap interval |
| ar2TapInter | Autoregressive coefficient of inter-tap intervals (characterizes relationship between inter-tap intervals at lag = 2) |
| meanDriftLeft | Mean of drift of finger position between consecutive taps in the left button |
| kurDriftLeft | Kurtosis of drift of finger position between consecutive taps in the left button |
| skewDriftRight | Skewness of drift of finger position between consecutive taps in the right button |
| madDriftRight | Median absolute deviation of drift of finger position between consecutive taps in the right button |

**(b) Voice**

| Feature | Definition |
|---|---|
| Shimmer->F0_dif_percent | Mean absolute difference of shimmer for successive cycles |
| Shimmer->F0_PQ11_classical_Baken | Classical Baken of the shimmer signal using 11 cycle samples |
| Shimmer->F0_abs0th_perturb | Zeroth order perturbation of shimmer signal |
| VFER->SNR_TKEO | Vocal fold excitation ratio of the signal-to-noise ratio of the Teager-Kaiser energy operator |
| std_MFCC_6th | Standard deviation of 6th Mel Frequency Cepstral Coefficient |
| det_entropy_log_4_coef | Log entropy of 4th detail coefficient |
| app_entropy_shannon_6_coef | Shannon entropy of 6th approximation coefficient |
| det_LT_entropy_log_4_coef | Log entropy of 4th detail coefficient (with prior F0 transform) |

Table 16 (continued)

**(c) Tapping + Voice**

| | Feature | Definition |
|---|---|---|
| **Tapping** | kurTapInter | Kurtosis of inter-tap interval |
| | madDriftRight | Median absolute deviation of drift of finger position between consecutive taps in the right button |
| **Voice** | Shimmer->F0_PQ5_classical_Schoentgen | Classical Schoentgen of the shimmer signal using 5 cycle samples. |
| | Shimmer->F0_FM | Frequency modulation of the shimmer signal |
| | mean_MFCC_1st | Mean of 1st Mel Frequency Cepstral Coefficient |
| | mean_MFCC_6th | Mean of 6th Mel Frequency Cepstral Coefficient |
| | std_8th_delta_delta | Standard deviation of the 8th delta delta (2nd derivative) Mel Frequency Cepstral Coefficient |
| | det_TKEO_mean_10_coef | Mean Teager-Kaiser energy operator 10th detail coefficient |
| | app_LT_entropy_log_1_coef | Log entropy of 1st approximation coefficient (with prior F0 transform) |
| | app_LT_entropy_log_5_coef | Log entropy of 5th approximation coefficient (with prior F0 transform) |

## 4.6 Conclusion

In this chapter, I developed s2SSL, a semi-supervised regression technique that applies both feature and instance selection to improve model building of datasets with few labeled instances and many available features. The model was applied to data collected from a smartphone app that performs telemonitoring of Parkinson's Disease patients. s2SSL utilized a particle swarm optimization method for selection of the smartphone-based telemonitoring features and a graph-sampling technique to reduce the computational time of training the semi-supervised learning algorithm. A high accuracy of 0.828 was achieving using both tapping and voice features collected from the smartphone app. Clinically relevant features were also selected and provide more information about which features are more effective at predicting the MDS-UPDRS Parkinson's Disease severity

score. s2SSL is capable of balancing data inclusivity (through SSL) and usability (through feature and sample selection). Future work will entail expanding the application of this model to other domains that have a large number of features and few labeled data instances. Such domains include telemonitoring other disease conditions such as Alzheimer's Disease, as well medical imaging of disease conditions, such as glioblastoma brain cancer or migraine.

# CHAPTER 5

# CONCLUSIONS AND FUTURE WORK

In the dawn of the information age in healthcare, improved technologies in imaging and telemonitoring have provided an unprecedented opportunity to harness massive amounts of data for improving patient care. Some data can be easier to acquire than others (for example biopsies of brain tumor are more difficult to obtain than images of brain tumor). Therefore, there is a need to develop predictive models that can utilize the easy-to-obtain data with the purpose of minimizing the need to get hard-to-obtain data in the future. At the same time, it is also important to choose only the most relevant information for model building to improve generalization capabilities on new patients. In my dissertation, I focused on building semi-supervised learning (SSL) models to balance data inclusivity and usability on a number of healthcare applications.

In my first topic, I developed an algorithm that integrates mechanistic models and machine learning to develop a new SSL model, which was applied to predict intra-tumor cell density of glioblastoma brain cancer using multiparametric MRI. This model was derived from imaging information of the patient and utilized scientific knowledge of glioblastoma diffusion and growth to make its prediction. Information from these mechanistic models was utilized as a sort of prior knowledge to guide the prediction of the machine learning model. The next topic focused on the development of a new constrained particle swarm optimization (cPSO)-based feature selection algorithm, which was applied to radiomics of glioblastoma brain cancer and migraine imaging. The algorithm was developed in a supervised learning setting and prepared to be further extended to an SSL setting. The final topic presented a novel SSL model with cPSO-integrated feature

selection and graph-based instance selection, which was applied to smartphone-based telemonitoring of Parkinson's Disease patients. I introduced a first-of-its-kind semi-supervised feature selection algorithm for regression tasks that combines cPSO feature selection and a graph-based instance selection method that reduces the model training time.

For future work, I plan to develop multi-task learning algorithms that take into account patient demographics (e.g., sex, genetic predisposition, etc.) during model training. Having models segmented to different patient types will help improve patient personalization and result in more accurate predictive models. Additionally, I plan to extend the semi-supervised models presented in this work to make recommendations on which hard-to-obtain data to collect in order to improve model prediction. This machine learning sub-field, known as active learning, can make recommendations on which samples to collect in real-time to improve the intelligence of data collection, resulting in improved model predictions. Having models work interactively with the clinicians will take healthcare data science to another level, allowing clinicians to directly interface with the models and obtain immediate recommendations on patient treatment and therapy.

# REFERENCES


Aghi M, Gaviani P, Henson JW, Batchelor TT, Louis DN, Barker FG. Magnetic resonance imaging characteristics predict epidermal growth factor receptor amplification status in glioblastoma. Clinical Cancer Research. 2005 Dec 15;11(24):8600-5.

Ang, J. C., Haron, H., & Hamed, H. N. A. (2015, June). Semi-supervised SVM-based feature selection for cancer classification using microarray gene expression data. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 468-477). Springer, Cham.

Arnold M. Headache Classification Committee of the International Headache Society (IHS) The International Classification of Headache Disorders. Cephalalgia. 2018;38(1):1-211.

Arora, S., Venkataraman, V., Zhan, A., Donohue, S., Biglan, K.M., Dorsey, E.R. and Little, M.A., (2015). Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study. Parkinsonism & related disorders, 21(6), pp.650-653.

Bache, K., & Lichman, M. (2013). UCI machine learning repository. Available: archive.ics.uci.edu/ml.

Baldock A, Ahn S, Rockne R, Neal M, et al. Patient-specific Metrics of Invasiveness Reveal Significant Prognostic Benefit of Resection in a Predictable Subset of Gliomas. PLoS One. 2014a;9(10).

Baldock AL, Rockne R, Boone A, Neal M, Mrugala MM, Rockhill JK, and Swanson KR. From Patient-Specific Mathematical Neuro-Oncology Towards Precision Medicine. Frontiers in Molecular and Cellular Oncology. 2013:3(62).

Baldock AL, Yagle K, Born DE, Ahn S, et al. Invasion and proliferation kinetics in enhancing gliomas predict IDH1 mutation status. Neuro Oncol. 2014b;16(6):779-86.

Barajas Jr RF, Hodgson JG, Chang JS, Vandenberg SR, Yeh RF, Parsa AT, McDermott MW, Berger MS, Dillon WP, Cha S. Glioblastoma multiforme regional genetic and cellular expression patterns: influence on anatomic and physiologic MR imaging. Radiology. 2010 Jan 6;254(2):564-76.

Barajas Jr RF, Phillips JJ, Parvataneni R, Molinaro A, Essock-Burns E, Bourne G, Parsa AT, Aghi MK, McDermott MW, Berger MS, Cha S. Regional variation in histopathologic features of tumor specimens from treatment-naive glioblastoma correlates with anatomic and physiologic MR Imaging. Neuro-oncology. 2012 Jul 1;14(7):942-54.

Barkia, H., Elghazel, H., & Aussem, A. (2011, December). Semi-supervised feature importance evaluation with ensemble learning. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on* (pp. 31-40). IEEE.

Beckmann CF, Smith SM. Tensorial extensions of independent component analysis for multisubject FMRI analysis. Neuroimage. 2005 Mar 1;25(1):294-311.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, *7*(Nov), 2399-2434.

Bellal, F., Elghazel, H., & Aussem, A. (2012). A semi-supervised feature ranking method with ensemble learning. *Pattern Recognition Letters*, *33*(10), 1426-1433.

Benabdeslem, K., & Hindawi, M. (2011, September). Constrained laplacian score for semi-supervised feature selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 204-218). Springer, Berlin, Heidelberg.

Bonavia R, Cavenee WK, Furnari FB. Heterogeneity maintenance in glioblastoma: a social network. Cancer research. 2011 Jun 15;71(12):4055-60.

Bonyadi MR, Michalewicz Z. Particle swarm optimization for single objective continuous space problems: a review. 2017.

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., ... & Friend, S. H. (2016). The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific data*, *3*, 160011.

Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R. The somatic genomic landscape of glioblastoma. Cell. 2013 Oct 10;155(2):462-77.

Brocks D, Assenov Y, Minner S, Bogatyrova O, Simon R and Koop C. Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. Cell reports. 2014;8(3):798-806.

Brown R, Zlatescu M, Sijben A, Roldan G, Easaw J, Forsyth P, Parney I, Sevick R, Yan E, Demetrick D, Schiff D. The use of magnetic resonance imaging to noninvasively detect genetic signatures in oligodendroglioma. Clinical Cancer Research. 2008 Apr 15;14(8):2357-62.

Calhoun VD, Adali T, Giuliani NR, Pekar JJ, Kiehl KA, Pearlson GD. Method for multimodal analysis of independent source differences in schizophrenia: combining gray matter structural and auditory oddball functional data. Human brain mapping. 2006 Jan;27(1):47-62.

Calhoun VD, Adali T. Feature-based fusion of medical imaging data. IEEE Transactions on Information Technology in Biomedicine. 2009 Sep;13(5):711-20.

Chan, P.L.S. and Holford, N.H.G., (2001). Drug treatment effects on disease progression. Annual review of pharmacology and toxicology, 41(1), pp.625-659.

Chang PD, Malone HR, Bowden SG, Chow DS, Gill BJ, Ung TH, Samanamud J, Englander ZK, Sonabend AM, Sheth SA and McKhann GM. A Multiparametric Model for Mapping Cellularity in Glioblastoma Using Radiographically Localized Biopsies. American Journal of Neuroradiology. 2017;38(5):890-8.

Chattopadhyay, R., Sun, Q., Fan, W., Davidson, I., Panchanathan, S., & Ye, J. (2012). Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *6*(4), 18.

Chen, L., Huang, R., & Huang, W. (2010, November). Graph-based semi-supervised weighted band selection for classification of hyperspectral data. In *Audio Language and Image Processing (ICALIP), 2010 International Conference on* (pp. 1123-1126). IEEE.

Cheng, H., Deng, W., Fu, C., Wang, Y., & Qin, Z. (2011). Graph-based semi-supervised feature selection with application to automatic spam image identification. In *Computer Science for Environmental Engineering and EcoInformatics* (pp. 259-264). Springer, Berlin, Heidelberg.

Cheng, Y., Cai, Y., Sun, Y., & Li, J. (2008, December). Semi-supervised feature selection under logistic i-relief framework. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on* (pp. 1-4). IEEE.

Chong CD, Gaw N, Fu Y, Li J, Wu T, Schwedt TJ. Migraine classification using magnetic resonance imaging resting-state functional connectivity data. Cephalalgia. 2017 Aug;37(9):828-44.

Chu X, Lu Q, Niu B, Wu T. Solving the distribution center location problem based on multi-swarm cooperative particle swarm optimizer. InInternational Conference on Intelligent Computing 2012 Jul 25 (pp. 626-633). Springer, Berlin, Heidelberg.

Correa NM, Eichele T, Adalı T, Li YO, Calhoun VD. Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI. Neuroimage. 2010 May 1;50(4):1438-45.

Correa NM, Li YO, Adali T, Calhoun VD. Fusion of fMRI, sMRI, and EEG data using canonical correlation analysis. InAcoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on 2009 Apr 19 (pp. 385-388). IEEE.

Corwin D, Holdsworth C, Rockne RC, Trister AD, et al. Toward patient-specific, biologically optimized radiation therapy plans for the treatment of glioblastoma. PLoS One. 2013;8.

Dai, K., Yu, H. Y., & Li, Q. (2013). A semisupervised feature selection with support vector machine. *Journal of Applied Mathematics*, *2013*.

Daniel O, Mauskop A. Nutraceuticals in acute and prophylactic treatment of migraine. Current treatment options in neurology. 2016 Apr 1;18(4):14.

Das S, Abraham A, Konar A. Swarm intelligence algorithms in bioinformatics. InComputational Intelligence in Bioinformatics 2008 (pp. 113-147). Springer, Berlin, Heidelberg.

Das S, Panigrahi BK, Pattnaik S. Nature-inspired algorithms for multi-objective optimization. Handbook of Research on Machine Learning Applications and Trends: Algorithms Methods and Techniques, Hershey, New York. 2009;1:95-108.

Dixon SJ, Brereton RG. Comparison of performance of five common classifiers represented as boundary methods: Euclidean distance to centroids, linear discriminant analysis, quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on data structure. Chemometrics and Intelligent Laboratory Systems. 2009 Jan 15;95(1):1-7.

Doquire, G., & Verleysen, M. (2013). A graph Laplacian based approach to semi-supervised feature selection for regression problems. *Neurocomputing*, *121*, 5-13.

Drabycz S, Roldán G, De Robles P, Adler D, McIntyre JB, Magliocco AM, Cairncross JG, Mitchell JR. An analysis of image texture, tumor location, and MGMT promoter methylation in glioblastoma using magnetic resonance imaging. Neuroimage. 2010 Jan 15;49(2):1398-405.

Durst CR, Raghavan P, Shaffrey ME, Schiff D, Lopes MB and Sheehan JP. Multimodal MR imaging model to predict tumor infiltration in patients with gliomas. Neuroradiology. 2014:56(2):107-15.

Eberhart RC, Shi Y, Kennedy J. Swarm Intelligence (Morgan Kaufmann series in evolutionary computation). Morgan Kaufmann Publishers; 2001.

Ene CI, Fine HA. Many tumors in one: a daunting therapeutic prospect. Cancer cell. 2011 Dec 13;20(6):695-7.

Ester M, Kriegel HP, Sander J and Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996.

Fan Y, Resnick SM, Wu X, Davatzikos C. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. Neuroimage. 2008 Jun 1;41(2):277-85.

Floreano D, Dürr P, Mattiussi C. Neuroevolution: from architectures to learning. Evolutionary Intelligence. 2008 Mar 1;1(1):47-62.

Fraser A, Burnell D. Computer models in genetics. Computer models in genetics. 1970.

Fujino A, Ueda N and Saito K. A hybrid generative/discriminative approach to semi-supervised classifier design. AAAI. 2005.

Gaw, N., Schwedt, T. J., Chong, C. D., Wu, T., & Li, J. (2018). A clinical decision support system using multi-modality imaging data for disease diagnosis. *IISE Transactions on Healthcare Systems Engineering*, *8*(1), 36-46.

Goetz, C. G., Stebbins, G. T., Wolff, D., DeLeeuw, W., Bronte-Stewart, H., Elble, R., ... & Wu, A. D. (2009). Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device. *Movement Disorders*, *24*(4), 551-556.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., ... & Dubois, B. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society*, *23*(15), 2129-2170.

Goldberg, A., Zhu, X., Singh, A., Xu, Z., & Nowak, R. (2009, April). Multi-manifold semi-supervised learning. In *Artificial Intelligence and Statistics* (pp. 169-176).

Grigorescu SE, Petkov N, Kruizinga P. Comparison of texture features based on Gabor filters. IEEE Transactions on Image processing. 2002 Oct;11(10):1160-7.

Groves AR, Beckmann CF, Smith SM, Woolrich MW. Linked independent component analysis for multimodal data fusion. Neuroimage. 2011 Feb 1;54(3):2198-217.

Gupta A, Young RJ, Shah AD, Schweitzer AD, Graber JJ, Shi W, Zhang Z, Huse J, Omuro AM. Pretreatment dynamic susceptibility contrast MRI perfusion in glioblastoma: prediction of EGFR gene amplification. Clinical neuroradiology. 2015 Jun 1;25(2):143-50.

Gutman DA, Cooper LA, Hwang SN, Holder CA, Gao J, Aurora TD, Dunn Jr WD, Scarpace L, Mikkelsen T, Jain R, Wintermark M. MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. Radiology. 2013 May;267(2):560-9.

Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research. 2003;3(Mar):1157-82.

Hajnal JV, Hill DL. Medical image registration. CRC press; 2001 Jun 27.

Han, Y., Park, K., & Lee, Y. K. (2011, August). Confident wrapper-type semi-supervised feature selection using an ensemble classifier. In *Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on* (pp. 4581-4586). IEEE.

Han, Y., Yang, Y., Yan, Y., Ma, Z., Sebe, N., & Zhou, X. (2015). Semisupervised feature selection via spline regression for video semantic recognition. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(2), 252-264.

Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983 Sep;148(3):839-43.

Haralick RM, Shanmugam K. Textural features for image classification. IEEE Transactions on systems, man, and cybernetics. 1973 Nov(6):610-21.

Hertrich, I., & Ackermann, H. (1995). Gender-specific vocal dysfunctions in Parkinson's disease: electroglottographic and acoustic analyses. *Annals of Otology, Rhinology & Laryngology*, *104*(3), 197-202.

Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L., & Gates, S. (1999). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural neurology*, *11*(3), 131-137.

Holmes, R. J., M. Oates, J., J. Phyland, D., & J. Hughes, A. (2000). Voice characteristics in the progression of Parkinson's disease. *International Journal of Language & Communication Disorders*, *35*(3), 407-418.

Holub, A., Welling, M., & Perona, P. (2005, December). Exploiting unlabelled data for hybrid object classification. In *Proc. Neural Information Processing Systems, Workshop Inter-Class Transfer* (Vol. 7, p. 2).

Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of educational psychology. 1933 Sep;24(6):417.

Hu LS, Eschbacher JM, Dueck AC, Heiserman JE, Liu S, Karis JP, Smith KA, Shapiro WR, Pinnaduwage DS, Coons SW, Nakaji P. Correlations between perfusion MR imaging cerebral blood volume, microvessel quantification, and clinical outcome using stereotactic analysis in recurrent high-grade glioma. American Journal of Neuroradiology. 2012a Jan 1;33(1):69-76.

Hu LS, Eschbacher JM, Heiserman JE, Dueck AC, Shapiro WR and Liu S. Reevaluating the imaging definition of tumor progression: perfusion MRI quantifies recurrent glioblastoma tumor fraction, pseudoprogression, and radiation necrosis to predict survival. Neuro-oncology. 2012b;14(7):919-30.

Hu LS, Ning S, Eschbacher JM, Baxter LC, et al. Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. Neuro-oncology. 2016;19(1):128-37.

Hu LS, Ning S, Eschbacher JM, Gaw N, et al. Multi-parametric MRI and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma. PLoS One. 2015;10(11).

Huang S, Li J, Ye J, Wu T, Chen K, Fleisher A, Reiman E. Identifying Alzheimer's disease-related brain regions from multi-modality neuroimaging data using sparse composite linear discrimination analysis. InAdvances in neural information processing systems 2011 (pp. 1431-1439).

Inda M, Bonavia R and Seoane J. Glioblastoma multiforme: a look inside its heterogeneous nature. Cancers. 2014;6(1):226-39.

Itakura H, Achrol AS, Mitchell LA, Loya JJ, Liu T, Westbroek EM, Feroze AH, Rodriguez S, Echegaray S, Azad TD, Yeom KW. Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. Science translational medicine. 2015 Sep 2;7(303):303ra138-.

Jackson P, Juliano J, Hawkins-Daarud AD, Rockne R, and Swanson KR. Patient-specific Mathematical Neuro-Oncology: Using a Simple Proliferation and Invasion Tumor Model to Inform Clinical Practice. Bulletin of Mathematical Biology. 2015 Mar 21;77(5):846-56.

Jain R, Poisson LM, Gutman D, Scarpace L, Hwang SN, Holder CA, Wintermark M, Rao A, Colen RR, Kirby J, Freymann J. Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor. Radiology. 2014 Mar 17;272(2):484-93.

Jarboui B, Cheikh M, Siarry P, Rebai A. Combinatorial particle swarm optimization (CPSO) for partitional clustering problem. Applied Mathematics and Computation. 2007 Sep 15;192(2):337-45.

Jarboui B, Damak N, Siarry P, Rebai A. A combinatorial particle swarm optimization for solving multi-mode resource-constrained project scheduling problems. Applied Mathematics and Computation. 2008 Jan 15;195(1):299-308.

Jiang, J., Lin, E., Wang, J., & Hanson, D. G. (1999). Glottographic measures before and after levodopa treatment in Parkinson's disease. *The Laryngoscope*, *109*(8), 1287-1294.

Jin C, Yuan K, Zhao L, Zhao L, Yu D, von Deneen KM, Zhang M, Qin W, Sun W, Tian J. Structural and functional abnormalities in migraine patients without aura. NMR in Biomedicine. 2013 Jan;26(1):58-64.

Kalakech, M., Biela, P., Macaire, L., & Hamad, D. (2011). Constraint scores for semi-supervised feature selection: A comparative study. *Pattern Recognition Letters*, *32*(5), 656-665.

Kassavetis, P., Saifee, T.A., Roussos, G., Drougkas, L., Kojovic, M., Rothwell, J.C., Edwards, M.J. and Bhatia, K.P., (2016). Developing a tool for remote digital assessment of Parkinson's Disease. Movement Disorders Clinical Practice, 3(1), pp.59-64.

Kennedy J, Eberhart RC. A discrete binary version of the particle swarm algorithm. InSystems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on 1997 Oct 12 (Vol. 5, pp. 4104-4108). IEEE.

Kettenring JR. Canonical analysis of several sets of variables. Biometrika. 1971 Dec 1;58(3):433-51.

Khan ME, Khan F. A comparative study of white box, black box and grey box testing techniques. Int. J. Adv. Comput. Sci. Appl. 2012 Jun;3(6).

Koza JR. Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems. Stanford, CA: Stanford University, Department of Computer Science; 1990 Jun 1.

Lainscsek, C., Rowat, P., Schettino, L., Lee, D., Song, D., Letellier, C. and Poizner, H., (2012). Finger tapping movements of Parkinson's disease patients automatically rated using nonlinear delay differential equations. Chaos: An Interdisciplinary Journal of Nonlinear Science, 22(1), p.013119.

Land, A. H., & Doig, A. G. (1960). An automatic method of solving discrete programming problems. *Econometrica: Journal of the Econometric Society*, 497-520.

Lawrence, N. D., & Jordan, M. I. (2005). Semi-supervised learning via Gaussian processes. In *Advances in neural information processing systems* (pp. 753-760).

Lee, C.Y., Kang, S.J., Hong, S.K., Ma, H.I., Lee, U. and Kim, Y.J., (2016). A validation study of a smartphone-based finger tapping application for quantitative assessment of bradykinesia in Parkinson's disease. PloS one, 11(7), p.e0158852.

Li, Y., Guan, C., Li, H., & Chin, Z. (2008). A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system. *Pattern Recognition Letters*, *29*(9), 1285-1294.

Lin GC, Wang WJ, Wang CM, Sun SY. Automated classification of multi-spectral MR images using linear discriminant analysis. Computerized Medical Imaging and Graphics. 2010 Jun 1;34(4):251-68.

Liu J, Pearlson G, Windemuth A, Ruano G, Perrone-Bizzozero NI, Calhoun V. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. Human brain mapping. 2009 Jan;30(1):241-55.

Liu, Y., Nie, F. P., Wu, J. G., & Chen, L. H. (2010, December). Semi-supervised feature selection based on label propagation and subset selection. In *Proceedings of the International Conference on Computer and Information Application*.

Liu, Y., Nie, F., Wu, J., & Chen, L. (2013). Efficient semi-supervised feature selection with noise insensitive trace ratio criterion. *Neurocomputing*, *105*, 12-18.

Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The design of SimpleITK. Frontiers in neuroinformatics. 2013 Dec 30;7:45.

Lu Z, Gao X, Wang L, Wen JR, Huang S. Noise-Robust Semi-Supervised Learning by Large-Scale Sparse Coding. InAAAI 2015 Jan 25 (pp. 2828-2834).

Lu, Z., & Wang, L. (2015). Noise-robust semi-supervised learning via fast sparse coding. *Pattern Recognition*, *48*(2), 605-612.

Lu, Z., Gao, X., Wang, L., Wen, J. R., & Huang, S. (2015, January). Noise-Robust Semi-Supervised Learning by Large-Scale Sparse Coding. In *AAAI* (pp. 2828-2834).

Ma, Z., Nie, F., Yang, Y., Uijlings, J. R., Sebe, N., & Hauptmann, A. G. (2012). Discriminating joint feature analysis for multimedia data understanding. *IEEE Transactions on Multimedia*, *14*(6), 1662-1672.

Ma, Z., Yang, Y., Nie, F., Uijlings, J., & Sebe, N. (2011, November). Exploiting the entire feature space with sparsity for automatic image annotation. In *Proceedings of the 19th ACM international conference on Multimedia* (pp. 283-292). ACM.

Mainero C, Boshyan J, Hadjikhani N. Altered functional magnetic resonance imaging resting-state connectivity in periaqueductal gray networks in migraine. Annals of neurology. 2011 Nov;70(5):838-45.

Maintz JA, Viergever MA. A survey of medical image registration. Medical image analysis. 1998 Mar 1;2(1):1-36.

Martelotto LG, Ng CK, Piscuoglio S, Weigelt B and Reis-Filho JS. Breast cancer intra-tumor heterogeneity. Breast Cancer Research. 2014;16(3):210.

Martınez-Montes E, Valdés-Sosa PA, Miwakeichi F, Goldman RI, Cohen MS. Concurrent EEG/fMRI analysis by multiway partial least squares. NeuroImage. 2004 Jul 1;22(3):1023-34.

Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer?. Nature Reviews Cancer. 2012 May 1;12(5):323-34.

McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ and Mastrogianakis GM. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061-8.

Mitchell JR, Jones C, Karlik SJ, Kennedy K, Lee DH, Rutt B, Fenster A. MR multispectral analysis of multiple sclerosis lesions. Journal of Magnetic Resonance Imaging. 1997 May 1;7(3):499-511.

Neal ML, Trister AD, Ahn S, Baldock A, et al. Response classification based on a minimal model of glioblastoma growth is prognostic for clinical outcomes and distinguishes progression from pseudoprogression. Cancer Res. 2013a;73(10):2976-86.

Neal ML, Trister AD, Cloke T, Sodt R, et al. Discriminating Survival Outcomes in Patients with Glioblastoma Using a Simulation-Based, Patient-Specific Response Metric. PLoS One. 2013b;8.

Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence. 2002 Jul;24(7):971-87.

Palermo S, Benedetti F, Costa T, Amanzio M. Pain anticipation: An activation likelihood estimation meta-analysis of brain imaging studies. Human brain mapping. 2015 May;36(5):1648-61.

Parsopoulos KE, editor. Particle swarm optimization and intelligence: advances and applications: advances and applications. IGI global; 2010 Jan 31.

Poli R, Kennedy J, Blackwell T. Particle swarm optimization. Swarm intelligence. 2007 Jun 1;1(1):33-57.

Pope WB, Chen JH, Dong J, Carlson MR, Perlina A, Cloughesy TF, Liau LM, Mischel PS, Nghiemphu P, Lai A, Nelson SF. Relationship between gene expression and enhancement in glioblastoma multiforme: exploratory DNA microarray analysis. Radiology. 2008 Oct;249(1):268-77.

Quinzán, I., Sotoca, J. M., & Pla, F. (2009, November). Clustering-based feature selection in semi-supervised problems. In *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on* (pp. 535-540). IEEE.

Ramig, L. A., Titze, I. R., Scherer, R. C., & Ringel, S. P. (1988). Acoustic analysis of voices of patients with neurologic disease: rationale and preliminary data. *Annals of Otology, Rhinology & Laryngology*, *97*(2), 164-172.

Ren, J., Qiu, Z., Fan, W., Cheng, H., & Philip, S. Y. (2008, May). Forward semi-supervised feature selection. In *Pacific-Asia conference on knowledge discovery and data mining*(pp. 970-976). Springer, Berlin, Heidelberg.

Roalf, D. R., Rupert, P., Mechanic-Hamilton, D., Brennan, L., Duda, J. E., Weintraub, D., ... & Moberg, P. J. (2018). Quantitative assessment of finger tapping characteristics in mild cognitive impairment, Alzheimer's disease, and Parkinson's disease. *Journal of neurology*, *265*(6), 1365-1375.

Robnik-Sikonja M and Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learning. 2003;53(1-2):23-69.

Rocca MA, Ceccarelli A, Falini A, Colombo B, Tortorella P, Bernasconi L, Comi G, Scotti G, Filippi M. Brain gray matter changes in migraine patients with T2-visible lesions: a 3-T MRI study. Stroke. 2006 Jul 1;37(7):1765-70.

Rockne R, Rockhill JK, Mrugala M, Spence AM, et al. Predicting the efficacy of radiotherapy in individual glioblastoma patients in vivo: a mathematical modeling approach. Phys Med Biol. 2010;55:3271-85.

Russo A, Tessitore A, Giordano A, Corbo D, Marcuccio L, De Stefano M, Salemi F, Conforti R, Esposito F, Tedeschi G. Executive resting-state network connectivity in migraine without aura. Cephalalgia. 2012 Oct;32(14):1041-8.

Ryoo I, Choi SH, Kim JH, Sohn CH, Kim SC, Shin HS, Yeom JA, Jung SC, Lee AL, Yun TJ, Park CK. Cerebral blood volume calculated by dynamic susceptibility contrast-enhanced perfusion MR imaging: preliminary correlation study with glioblastoma genetic profiles. PloS one. 2013 Aug 19;8(8):e71704.

Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. bioinformatics. 2007 Oct 1;23(19):2507-17.

Scholkopf B, Herbrich R and Smola A. A generalized representer theorem. Computational learning theory. 2001;416-426.

Schüpbach, M.W., Corvol, J.C., Czernecki, V., Djebara, M.B., Golmard, J.L., Agid, Y. and Hartmann, A., (2009). The segmental progression of early untreated Parkinson disease: a novel approach to clinical rating. Journal of Neurology, Neurosurgery & Psychiatry.

Schwedt TJ, Chong CD, Chiang CC, Baxter L, Schlaggar BL, Dodick DW. Enhanced pain-induced activity of pain-processing regions in a case-control study of episodic migraine. Cephalalgia. 2014 Oct;34(12):947-58.

Schwedt TJ, Larson-Prior L, Coalson RS, Nolan T, Mar S, Ances BM, Benzinger T, Schlaggar BL. Allodynia and descending pain modulation in migraine: a resting state functional connectivity analysis. Pain medicine. 2014 Jan 1;15(1):154-65.

Schwedt TJ, Schlaggar BL, Mar S, Nolan T, Coalson RS, Nardos B, Benzinger T, Larson-Prior LJ. Atypical resting-state functional connectivity of affective pain regions in chronic migraine. Headache: The Journal of Head and Face Pain. 2013 May;53(5):737-51.

Schwedt TJ. Multisensory integration in migraine. Current opinion in neurology. 2013 Jun;26(3):248.

Schwedt, T. J., Chong, C. D., Wu, T., Gaw, N., Fu, Y., & Li, J. (2015). Accurate classification of chronic migraine via brain magnetic resonance imaging. *Headache: The Journal of Head and Face Pain*, *55*(6), 762-777.

Sethian JA. Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science. Cambridge university press; 1999 Jun 13.

Sheikhpour, R., Sarram, M. A., Gharaghani, S., & Chahooki, M. A. Z. (2017). A survey on semi-supervised feature selection methods. *Pattern Recognition*, *64*, 141-158.

Shi, C., Ruan, Q., & An, G. (2014). Sparse feature selection based on graph Laplacian for web image annotation. *Image and Vision Computing*, *32*(3), 189-201.

Simons LE, Moulton EA, Linnman C, Carpino E, Becerra L, Borsook D. The human amygdala and pain: evidence from neuroimaging. Human brain mapping. 2014 Feb;35(2):527-38.

Skodda, S., Rinsche, H., & Schlegel, U. (2009). Progression of dysprosody in Parkinson's disease over time—a longitudinal study. *Movement disorders: official journal of the Movement Disorder Society*, *24*(5), 716-722.

Sodt R, Rockne R, Neal ML, Kalet I, et al. Quantifying the role of antisotropic invasion in human glioblastoma. New York: Springer. 2010.

Song, X., Zhang, J., Han, Y., & Jiang, J. (2016). Semi-supervised feature selection via hierarchical regression for web image classification. *Multimedia Systems*, *22*(1), 41-49.

Sottoriva A, Spiteri I, Piccirillo SG, Touloumis A, Collins VP, Marioni JC, Curtis C, Watts C, Tavaré S. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proceedings of the National Academy of Sciences. 2013 Mar 5;110(10):4009-14.

Stadlbauer A, Ganslandt O, Buslei R, Hammen T, Gruber S, Moser E, Buchfelder M, Salomonowitz E, Nimsky C. Gliomas: histopathologic evaluation of changes in directionality and magnitude of water diffusion at diffusion-tensor MR imaging. Radiology. 2006 Sep;240(3):803-10.

Stankewitz A, May A. Increased limbic and brainstem activity during migraine attacks following olfactory stimulation. Neurology. 2011 Jul 20:WNL-0b013e318227e4a8.

Storn R, Price K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization. 1997 Dec 1;11(4):341-59.

Stupp R, Mason WP, Van Den Bent MJ, Weller M, Fisher B, Taphoorn MJ, Belanger K, Brandes AA, Marosi C, Bogdahn U and Curschmann J. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New England Journal of Medicine. 2005;352(10):987-86.

Sui J, Pearlson G, Caprihan A, Adali T, Kiehl KA, Liu J, Yamamoto J, Calhoun VD. Discriminating schizophrenia and bipolar disorder by fusing fMRI and DTI in a multimodal CCA+ joint ICA model. Neuroimage. 2011 Aug 1;57(3):839-55.

Sun, S., Hussain, Z., & Shawe-Taylor, J. (2014). Manifold-preserving graph reduction for sparse semi-supervised learning. *Neurocomputing*, *124*, 13-21.

Swanson KR, Alvord EC and Murray J. A quantitative model for differential motility of gliomas in grey and white matter. Cell Prolif. 2000;33(5):317-29.

Swanson KR, Alvord EC and Murray J. Virtual brain tumours (gliomas) enhance the reality of medical imaging and highlight inadequacies of current therapy. Br J Cancer. 2002;86(1):14-8.

Swanson KR, Alvord EC, Murray JD and Rockne RC. Method and system for characterizing tumors. United States of America Patent 8571844, 29 October 2013.

Swanson KR, Bridge C, Murray J and Alvord EC. Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion. J Neurol Sci. 2003;216(1):1-10.

Tang, B., & Zhang, L. (2018, August). Semi-supervised Feature Selection Based on Logistic I-RELIEF for Multi-classification. In *Pacific Rim International Conference on Artificial Intelligence* (pp. 719-731). Springer, Cham.

Tanha, J., van Someren, M., & Afsarmanesh, H. (2017). Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics*, *8*(1), 355-370.

Taylor Tavares, A.L., Jefferis, G.S., Koop, M., Hill, B.C., Hastie, T., Heit, G. and Bronte - Stewart, H.M., (2005). Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation. Movement disorders: official journal of the Movement Disorder Society, 20(10), pp.1286-1298.

Titze, I.R. (2000). Principals of Voice Production. National Center for Voice and Speech, Iowa City, US, 2nd edition.

Tsanas, A. (2012). *Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning* (Doctoral dissertation, Oxford University, UK).

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *Journal of the royal society interface*, *8*(59), 842-855.

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC. N4ITK: improved N3 bias correction. IEEE transactions on medical imaging. 2010 Jun;29(6):1310-20.

Tykocinski ES, Grant RA, Kapoor GS, Krejza J, Bohman LE, Gocke TA, Chawla S, Halpern CH, Lopinto J, Melhem ER, O'rourke DM. Use of magnetic perfusion-weighted imaging to determine epidermal growth factor receptor variant III expression in glioblastoma. Neuro-oncology. 2012 Apr 4;14(5):613-23.

Urish KL, Williams AA, Durkin JR, Chu CR, OAI Investigators Group. Registration of magnetic resonance image series for knee articular cartilage analysis: data from the osteoarthritis initiative. Cartilage. 2013 Jan;4(1):20-7.

Van Meter T, Dumur C, Hafez N, Garrett C, Fillmore H, Broaddus WC. Microarray analysis of MRI-defined tissue samples in glioblastoma reveals differences in regional expression of therapeutic targets. Diagnostic Molecular Pathology. 2006 Dec 1;15(4):195-205.

Wan, X. (2009, August). Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1* (pp. 235-243). Association for Computational Linguistics.

Wang CH, Rockhill JK, Mrugala M, Peacock DL, et al. Prognostic significance of growth kinetics in newly diagnosed glioblastomas revealed by combining serial imaging with a novel biomathematical model. Cancer Res. 2009;69(23):9133-40.

Wang X, Yang J, Teng X, Xia W, Jensen R. Feature selection based on rough sets and particle swarm optimization. Pattern recognition letters. 2007 Mar 1;28(4):459-71.

Wang, B., Jia, Y., & Yang, S. (2008, December). Forward semi-supervised feature selection based on Relevant set correlation. In *Computer Science and Software Engineering, 2008 International Conference on* (Vol. 4, pp. 210-213). IEEE.

Wang, F., & Zhang, C. (2008). Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering*, *20*(1), 55-67.

Xu L, Pearlson G, Calhoun VD. Joint source based morphometry identifies linked gray and white matter group differences. Neuroimage. 2009 Feb 1;44(3):777-89.

Xu, Z., King, I., Lyu, M. R. T., & Jin, R. (2010). Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural networks*, *21*(7), 1033-1047.

Yang D, Rao G, Martinez J, Veeraraghavan A, Rao A. Evaluation of tumor-derived MRI-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma. Medical physics. 2015 Nov 1;42(11):6725-35.

Yang H, Liu J, Sui J, Pearlson G, Calhoun VD. A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia. Frontiers in human neuroscience. 2010 Oct 25;4:192.

Yang, L., & Wang, L. (2007, August). Simultaneous feature selection and classification via semi-supervised models. In *Natural Computation, 2007. ICNC 2007. Third International Conference on* (Vol. 1, pp. 646-650). IEEE.

Yang, M., Chen, Y. J., & Ji, G. L. (2010, July). Semi_Fisher Score: A semi-supervised method for feature selection. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on* (Vol. 1, pp. 527-532). IEEE.

Yang, W., Hou, C., & Wu, Y. (2011, October). A Semi-supervised Method for Feature Selection. In *Computational and Information Sciences (ICCIS), 2011 International Conference on* (pp. 329-332). IEEE.

Yoon, H., & Li, J. (2019). A Novel Positive Transfer Learning Approach for Telemonitoring of Parkinson's Disease. *IEEE Transactions on Automation Science and Engineering*, *16*(1), 180-191.

Yu G, Liu Y, Thung KH, Shen D. Multi-task linear programming discriminant analysis for the identification of progressive MCI individuals. PloS one. 2014 May 12;9(5):e96458.

Yuan L, Wang Y, Thompson PM, Narayan VA, Ye J, Alzheimer's Disease Neuroimaging Initiative. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. NeuroImage. 2012 Jul 2;61(3):622-32.

Zacharaki EI, Wang S, Chawla S, Soo Yoo D, Wolf R, Melhem ER, Davatzikos C. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. Magnetic resonance in medicine. 2009 Dec 1;62(6):1609-18.

Zeng, Z., Wang, X., Zhang, J., & Wu, Q. (2016). Semi-supervised feature selection based on local discriminative information. *Neurocomputing*, *173*, 102-109.

Zhang D, Wang Y, Zhou L, Yuan H, Shen D, Alzheimer's Disease Neuroimaging Initiative. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage. 2011 Apr 1;55(3):856-67.

Zhang Q, Wu Q, Zhang J, He L, Huang J, Zhang J, Huang H, Gong Q. Discriminative analysis of migraine without aura: using functional and structural MRI with a multi-feature classification approach. PloS one. 2016 Sep 30;11(9):e0163875.

Zhang, K., Wang, Q., Lan, L., Sun, Y., & Marsic, I. (2014). Sparse semi-supervised learning on low-rank kernel. *Neurocomputing*, *129*, 265-272.

Zhao, J., Lu, K., & He, X. (2008). Locality sensitive semi-supervised feature selection. *Neurocomputing*, *71*(10-12), 1842-1849.

Zhao, M., Jiao, L., Feng, J., & Liu, T. (2014). A simplified low rank and sparse graph for semi-supervised learning. *Neurocomputing*, *140*, 84-96.

Zhao, Z., & Liu, H. (2007, April). Semi-supervised feature selection via spectral analysis. In *Proceedings of the 2007 SIAM international conference on data mining* (pp. 641-646). Society for Industrial and Applied Mathematics.

Zhou ZH, Zhan DC and Yang Q. Semi-supervised learning with very few labeled training examples. AAAI. 2007.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. In *Advances in neural information processing systems* (pp. 321-328).

Zhou, Z. H., Zhan, D. C., & Yang, Q. (2007, July). Semi-supervised learning with very few labeled training examples. In *AAAI* (pp. 675-680).

Zhu, X., & Lafferty, J. (2005, August). Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *Proceedings of the 22nd international conference on Machine learning* (pp. 1052-1059). ACM.

Zhu, X., Ghahramani, Z., & Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)* (pp. 912-919).

Zoski, K. W., & Jurs, S. (1996). An objective counterpart to the visual scree test for factor analysis: The standard error scree. *Educational and Psychological Measurement*, *56*(3), 443-451.

Zuo, L., Li, L., & Chen, C. (2015). The graph based semi-supervised algorithm with ℓ1-regularizer. *Neurocomputing*, *149*, 966-974.

# APPENDIX A

## SUPPLEMENTARY MATERIALS FOR CHAPTER 2

Proof of Theorem 1

Given any function $f \in \mathcal{H}_K$, $f$ can be uniquely comprised of $f_{\|}$ and $f_{\perp}$, where $f_{\|}$ is in the linear subspace spanned by the kernel functions $\{K(\mathbf{x}_i,\cdot)\}_{i=1}^{L+U}$ and $f_{\perp}$ is the orthogonal component. By the reproducing property, the value of $f$ on any point $\mathbf{x}_j, 1 \leq j \leq L+U$ is independent of $f_{\perp}$, as shown below:

$$f(\mathbf{x}_j) = \langle f, K(\mathbf{x}_j,\cdot)\rangle = \langle \textstyle\sum_{i=1}^{L+U} \alpha_i K(\mathbf{x}_i,\cdot), K(\mathbf{x}_j,\cdot)\rangle + \langle f_{\perp}, K(\mathbf{x}_j,\cdot)\rangle$$

It follows that $\langle K(\mathbf{x}_i,\cdot), K(\mathbf{x}_j,\cdot)\rangle = K(\mathbf{x}_i, \mathbf{x}_j)$ and $\langle f_{\perp}, K(\mathbf{x}_j,\cdot)\rangle$ vanishes. Therefore, the above formulation simplifies to

$$f(\mathbf{x}_j) = \textstyle\sum_{i=1}^{L+U} \alpha_i K(\mathbf{x}_i, \mathbf{x}_j),$$

which means that the terms of the optimization in (3) only rely on the gram matrix of the kernel function and the coefficients $\{\alpha_i\}_{i=1}^{L+U}$. Furthermore, the norm of $f$ in $\mathcal{H}_K$ has the following decomposition:

$$\|f\|_K^2 = \left\|\textstyle\sum_{i=1}^{L+U} \alpha_i K(\mathbf{x}_i,\cdot)\right\|_K^2 + \|f_{\perp}\|_K^2 \geq \left\|\textstyle\sum_{i=1}^{L+U} \alpha_i K(\mathbf{x}_i,\cdot)\right\|_K^2$$

The above inequality is true because $f_{\perp}$ will only increase $\|f\|_K^2$, so it follows that the minimizer of (3) must result in $f_{\perp} = 0$, leading to

$$f^*(\mathbf{x}) = \textstyle\sum_{i=1}^{L+U} \alpha_i K(\mathbf{x}_i, \mathbf{x}).$$ ■

Virtual biopsy selection procedure

Step 1: For each patient, count the number of biopsy samples with density >70%. Denote this number by $r$. $r'$ is the number of real biopsies with low-density. $v = r - r'$ is the number of virtual biopsy samples with low-density (<30%) that are to be found, in order to create balanced samples for the patient.

Step 2: Locate the BAT for the patient by subtracting the ROI segmented on T1+C from the ROI segmented on T2W. On the PI-estimated density map over the BAT, pick a sub-area to take virtual biopsy from according to the following biological criteria:

1) The sub-area needs to be away from the skull and the midline of the brain, since PI estimation tends to be less accurate at locations with physical barriers.
2) The sub-area should be close to the peripheral of the T2W ROI, where there is much lower chance to harbor high cell density.
3) Considering spatial continuity of cell density distribution, the PI estimation at a neighborhood of the biopsy sample should be more likely to be accurate if there is a real biopsy sample with low density whose PI density is also low. If the density of the real biopsy sample disagrees with PI density, the neighborhood of the sample should be avoided.

Step 3: On the sub-area that is picked according to *Step 2*, the following statistical criteria are further applied to select $v$ virtual biopsy samples:

1) Spatial consistency of PI density: For each pixel in the sub-area, place an 8x8 voxel box around it. Then, compute the mean and variance of PI densities over the 64 pixels within the box. Keep the boxes with a low mean (<30%) and a low variance

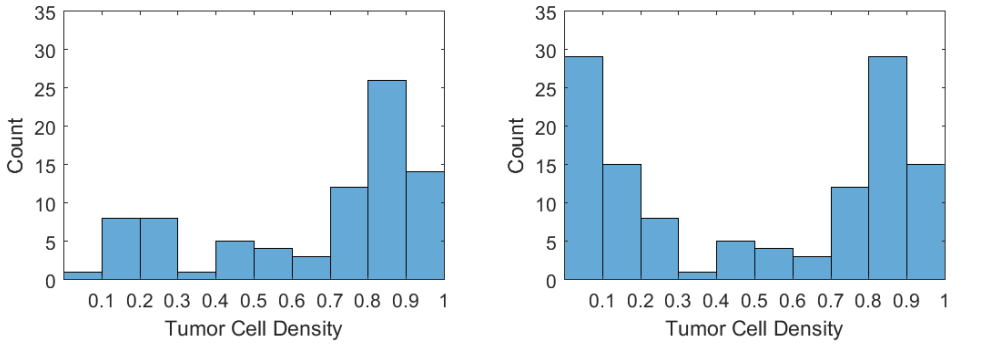
as potential virtual biopsy samples.

2) Separation in the imaging feature space: Good virtual biopsy samples need to be at a certain distance away from each other in the input (imaging features) space – called leverage samples in statistics – in order to stabilize model fitting. To find the leverage samples, I use a highly flexible and efficient clustering algorithm called DBSCAN (Ester et al. 1996) to cluster the boxes that have survived sub-step 1) using imaging features. Parameters of DBSCAN are set to produce approximately $v$ clusters. Then, one box from each cluster is picked as the virtual biopsy sample.

MRI protocol, image co-registration, and normalization

All imaging was performed on a 3 Tesla system (Sigma HDx; GE-Healthcare, Milwaukee, Wisconsin) within 1 day prior to stereotactic surgery. Conventional MRI included standard pre- and post-contrast T1-Weighted and pre-contrast T2-Weighted sequences. In addition, DTI imaging was performed using Spin-Echo Echo-planar imaging (EPI). I normalized the signal for T1+C, T2W, and EPI+C image datasets using the Simple Insight Segmentation and Registration Toolkit (SimpleITK v1.0.1) (Lowekamp et al. 2013) in Python (v3.6.2). The CurvatureFlow algorithm was applied to remove image noise (Sethian 1999) and the N4ITK algorithm to correct for image intensity nonuniformity bias that could be due to factors such as local magnetic field heterogeneity (Tustison et al. 2010). Following these corrections, the cerebrospinal fluid (CSF) of the lateral ventricles was used as a reference tissue to normalize the intensity distributions of each dataset using a previously described linear scaling process (Mitchell et al. 1997). Several parametric

maps were calculated, such as mean diffusivity (MD) and fractional anisotrophy (FA) based on previously published methods (Hu et al. 2015). Also, DSC-pMRI were acquired as previously described and calculated relative cerebral blood volume (rCBV) using IB Neuro (Hu et al. 2015). Multiparametric images were coregistered from each patient according to (Hu et al. 2012b), using tools from ITK (www.itk.org) and IB (Imaging Biometrics) Suite. After coregistration, the imaging data had a plane voxel resolution of ~1.2 mm (256 × 256 matrix) and slice thickness of 3 mm. Following our previous publications (Hu et al. 2015, Hu et al. 2016), six multiparametric images were included in the present study, including T1+C, T2W, EPI+C, MD, FA, and rCBV.




(a) Real biopsies (b) Real and virtual biopsies

Figure A1: Distribution of cell density in (a) real biopsies, and (b) virtual biopsies.

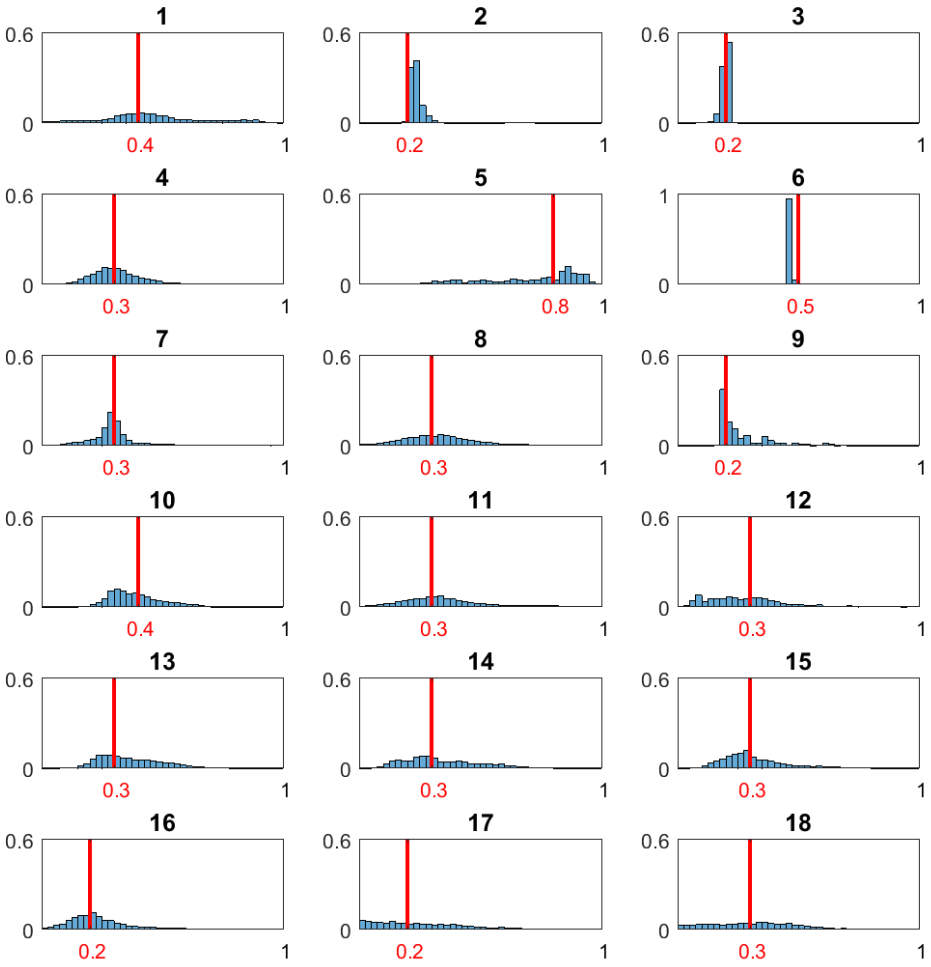



Figure A2: Histograms of ML-PI prediction at the BAT for each patient (1-18). The median of each histogram is indicated in red.

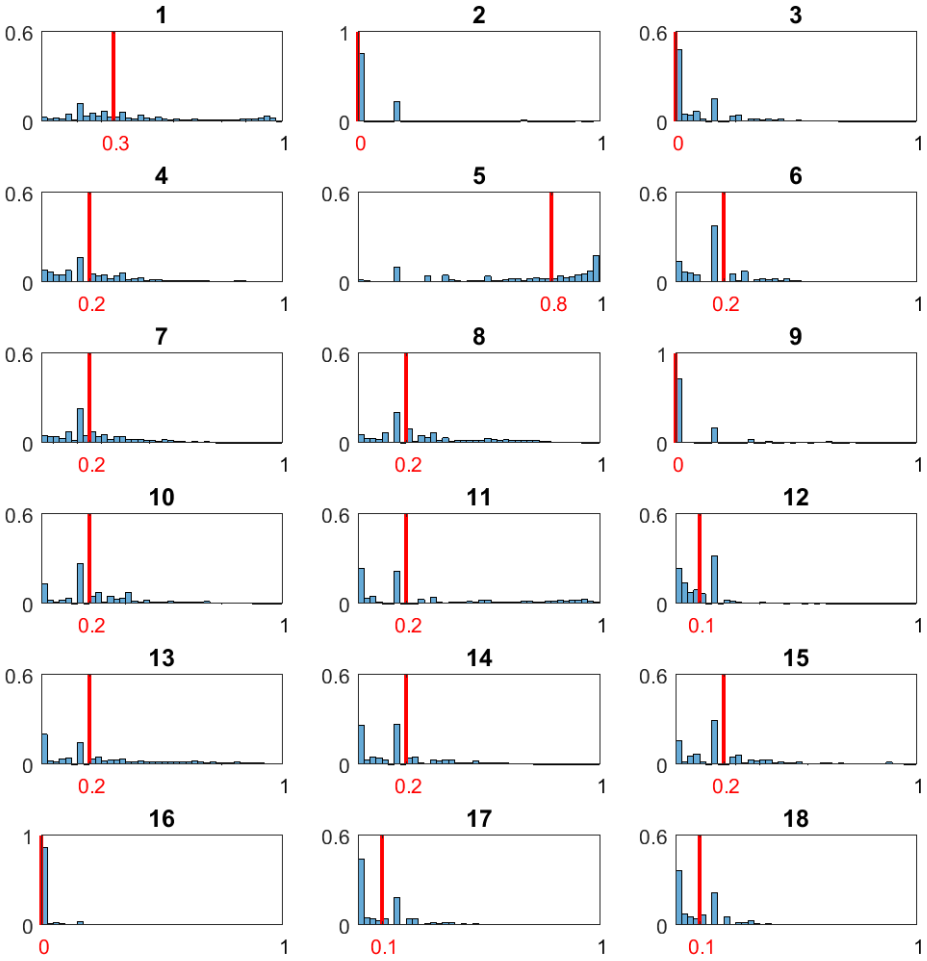



Figure A3: Histograms of PI prediction at the BAT for each patient (1-18). The median of each histogram is indicated in red.

Table A1: Patient-wise MAPEs of ML-PI, PI, and ML

| Patient | #biopsy samples | ML-PI | PI | ML |
|---|---|---|---|---|
| 1 | 7 | 0.086 ± 0.098* | 0.147 ± 0.133 | 0.255 ± 0.163 |
| 2 | 2 | 0 ± 0* | 0.204 ± 0.066 | 0.098 ± 0.034 |
| 3 | 5 | 0.085 ± 0.148* | 0.295 ± 0.158 | 0.229 ± 0.291 |
| 4 | 3 | 0.094 ± 0.039* | 0.157 ± 0.16 | 0.169 ± 0.122 |
| 5 | 2 | 0.006 ± 0.008* | 0.017 ± 0.024 | 0.219 ± 0.289 |
| 6 | 5 | 0.294 ± 0.041* | 0.682 ± 0.067 | 0.508 ± 0.145 |
| 7 | 3 | 0.106 ± 0.184* | 0.172 ± 0.253 | 0.424 ± 0.085 |
| 8 | 6 | 0.117 ± 0.066* | 0.203 ± 0.06 | 0.164 ± 0.121 |
| 9 | 3 | 0.166 ± 0.219* | 0.251 ± 0.337 | 0.24 ± 0.192 |
| 10 | 3 | 0.075 ± 0.09* | 0.144 ± 0.15 | 0.111 ± 0.089 |
| 11 | 6 | 0.044 ± 0.062* | 0.223 ± 0.211 | 0.103 ± 0.096 |
| 12 | 4 | 0.135 ± 0.155* | 0.307 ± 0.304 | 0.229 ± 0.259 |
| 13 | 14 | 0.164 ± 0.166* | 0.243 ± 0.224 | 0.193 ± 0.204 |
| 14 | 4 | 0.119 ± 0.119 | 0.21 ± 0.177 | 0.1 ± 0.124* |
| 15 | 3 | 0 ± 0* | 0.007 ± 0.011 | 0.002 ± 0.004 |
| 16 | 4 | 0.077 ± 0.059* | 0.084 ± 0.101 | 0.22 ± 0.091 |
| 17 | 2 | 0 ± 0* | 0.344 ± 0.224 | 0.158 ± 0.079 |
| 18 | 6 | 0.04 ± 0.043* | 0.169 ± 0.188 | 0.096 ± 0.114 |

The smallest mean absolute prediction error (MAPE) among the ML-PI, PI, and ML models is emphasized with an asterisk (*).

# APPENDIX B

## SUPPLEMENTARY MATERIALS FOR CHAPTER 4

‘make_s_curve’ Function Definition

```
make_s_curve(n_samples, noise, random_state)
```

Input:

`n_samples`**:** # sample points on the S curve

`noise`**:** standard deviation of the Gaussian noise

`random_state`**:** determines the random number generation for dataset creation; pass an integer for reproducible output across multiple function calls

Output:

$x = [x_1 | x_2]$: ‘the points’ along the S curve (array of size `n_samples` $\times 2$)

$t$: the univariate position of the sample according to the main dimension of the points in the manifold (array of size (`n_samples` $\times 1$)

Mathematical definitions:

$t = 3\pi \cdot Z_{n\_samples \times 1}; Z_i \sim U(-0.5, 0.5), i = 1, \dots, n\_samples$

$x_1 = \sin(t); x_2 = sign(t) \cdot (\cos(t) - 1)$; $sign(t)$ returns the sign of $t$: +1, -1, or 0

$x = [x_1 | x_2]$; define X as the concatenation of $X_1$ and $X_2$

$x = x + noise \cdot W_{n\_samples \times 2}; W_{ij} \sim N(0,1), i = 1, \dots, n\_samples, j = 1,2$