A New Algorithm for Shared Simultaneous Learning of Alzheimer's Disease Progression

by

Pushkar Mutha

B.E. Electronics Engineering, University of Mumbai, 2018

Submitted to the Graduate Faculty of the

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

Pushkar Mutha

It was defended on

July 13, 2021

and approved by

Liang Zhan, Ph.D., Assistant Professor,
Department of Electrical and Computer Engineering

Murat Akcakaya, Ph.D., Associate Professor,
Department of Electrical and Computer Engineering

Ahmed Dallal, Ph.D., Assistant Professor,
Department of Electrical and Computer Engineering

Thesis Advisor: Liang Zhan, Ph.D., Assistant Professor,
Department of Electrical and Computer Engineering

A New Algorithm for Shared Simultaneous Learning of Alzheimer's Disease Progression

Pushkar Mutha, MS

University of Pittsburgh, 2021

Alzheimer's Disease (AD), a progressive neurodegenerative disease, is the most common form of dementia in older adults. It is preceded by stages of subtle cognitive decline called as Mild Cognitive Impairment (MCI), which is further stratified into Early (EMCI) and Late (LMCI) stages. Several imaging biomarkers are being investigated for early and accurate diagnosis as well as prognosis, and traditional approaches have generally focused on training multiple independent binary classifiers for distinguishing between Normal Controls (NC), EMCI, LMCI and AD subjects. However, these multiple one vs one classifiers could hold complementary information and sharing this information during the training may improve predictive performance.

We introduce a new framework to perform Shared Simultaneous Learning (SSL) of sparse logistic regression classifiers for NC vs EMCI, EMCI vs LMCI, and LMCI vs AD classification. We achieve this by adding a new term to the logistic loss function to enforce the weight vectors to be similar to each other. We introduce a constraint to minimize the squared Euclidean distance between the three weight vectors. A smooth approximation for the absolute value function is used and the model is optimized using gradient descent with line search. For each classifier, at the current gradient descent step, weights from the other two classifiers are shared.

We evaluated this algorithm on Structural Brain Connectome Networks generated from diffusion MRI of 202 subjects from the multicenter Alzheimer's Disease Neuroimaging Initiative 2 (ADNI2) dataset. The normalized adjacency matrices were vectorized and passed as input for training along with the corresponding class labels. SSL outperformed independently trained

multiple linear binary classifiers and achieved an average AUC of 0.53 for NC vs EMCI, 0.68 for EMCI vs LMCI, and 0.73 for LMCI vs AD classification. We also analyzed the brain connectivity patterns associated with highest odds ratio and show that abnormal inter-hemispheric connectivity patterns are indicative of EMCI vs LMCI whereas the right hemisphere of the brain is involved in the later stages.

Table of Contents

List of Tables

List of Figures

Preface

I am extremely grateful to my adviser, Dr. Liang Zhan for his constant support and mentoring throughout my 2 years at the University of Pittsburgh. I also wish to thank Dr. Murat Akcakaya and Dr. Ahmed Dallal for agreeing to be on my thesis committee. I would like to extend my deepest gratitude to my colleagues at Biogen with whom I had an excellent learning experience during my co-op. I would also like to extend my sincere thanks to my undergraduate professors, Dr. Saurav Mitra, Prof. Sheetal Patil and Prof. Akhil Masurkar for I would not have been able to pursue graduate school in the US without their advice and encouragement. I am deeply indebted to my family and friends for their love, support and encouragement. Special thanks to my friends, Tushar, Saurabh, Sana, and Anisha for being with me through my highs and lows. Lastly, I am very grateful to all the faculty and staff at the University of Pittsburgh for the help and resources they provided during my graduate studies.

1.0 Introduction

Alzheimer's Disease (AD) is the most common form of dementia in people over 65. It is a progressive neurological disease which is associated with excessive β amyloid plaques and neurofibrillary tangles in the brain. Progressive neuronal death results in the decline of cognitive abilities and severely affects the quality of life of affected patients. The symptoms are mild at first but cause large scale tissue loss and neuronal atrophy with progression (Kocahan & Doğan, 2017). It is estimated that about 6.2 million Americans suffer from AD and the number is projected to rise to nearly 13.8 million by 2060. While the exact causes of AD are still unknown, it is believed that the combination of several risk factors such as increasing age, genetics, lifestyle and environmental factors are at play ("2021 Alzheimer's Disease Facts and Figures," 2021).

Mild Cognitive Impairment (MCI) is the prodromal stage of AD characterized by memory impairment beyond that expected for normal cognitive aging (Petersen et al., 1999). Patients with MCI are 28.9% to 33.6% more likely to develop AD (Mitchell & Shiri-Feshki, 2009). In the Alzheimer's Disease Neuroimaging Initiative 2 (ADNI2) protocol, MCI is further stratified into Early Mild Cognitive Impairment (EMCI) and Late Mild Cognitive Impairment (LMCI) stages based on education adjusted scores from the Logical Memory II subscale from the Wechsler Memory Scale. Recently, the FDA authorized the first ever disease modifying therapy, aducanumab (*FDA's Decision to Approve New Treatment for Alzheimer's Disease | FDA*, n.d.), for the treatment of Alzheimer's Disease. Administering pharmacological interventions at early stages could help slow the progression and/or lessen the severity of symptoms, ultimately improving patient's lives. There is thus an unmet need for accurate detection and prognostication for maximum patient benefit.

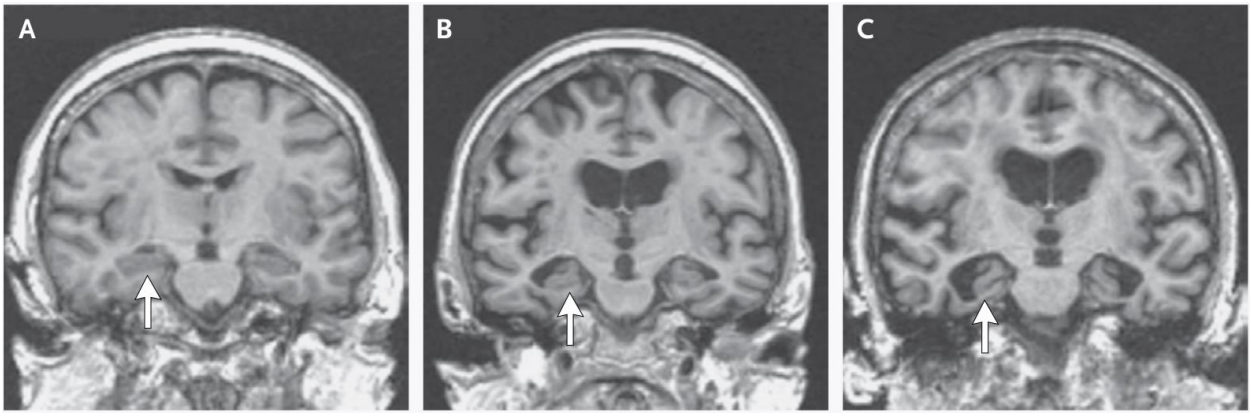

Figure 1: Coronal T1 MRI slice depicting the brain of a A) Normal, B) Mild Cognitive Impairment, C) Alzheimer's Disease subjects.

The white arrow points towards Hippocampal atrophy. (Petersen, 2011)

Reproduced with permission from Petersen, Ronald C. "Mild Cognitive Impairment." ***The New England Journal of Medicine*** **364.23 (2011): 2227–2234. Web., Copyright Massachusetts Medical Society.**

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique that can capture high resolution structural as well as functional 3D images of the brain. Figure 1 shows a slice of T1 weighted structural brain MRI of a A) Normal, B) Mild Cognitive Impairment, and C) Alzheimer's Disease subject (Petersen, 2011). The atrophy of the brain and specifically the hippocampus marked by the white arrow as well as the enlargement of ventricles is evident in these images. Several biomarkers for AD and its early stages are currently under investigation using automated analyses of neuroimaging technologies such as structural MRI, functional MRI, Positron Emission Tomography, etc. (Márquez & Yassa, 2019). Multimodal approaches have also been proposed that combine information from multiple modalities such as MRI scans, PET scans, clinical reports or scores from standard disease severity scales, etc. for diagnosis, prognosis, treatment evaluation or disease stratification (Chen et al., 2020; Chételat, 2018; Kim et al., 2020; Teipel et al., 2015).

Diffusion Weighted MRI (dMRI) is a neuroimaging technique based on the principle of diffusion of water molecules in human tissue. The axonal structure impedes random Brownian motion of water molecules present within them along the axonal orientation. This anisotropic diffusion of water molecules along the axonal orientation is exploited by dMRI methods, including Diffusion Tensor Imaging (DTI), to probe the anatomical connectivity between different brain regions. Tractography algorithms applied to these dMRIs can generate high resolution maps of anatomical brain connectivity (Tournier, 2019). Diffusion MRI is particularly useful in AD to assess the extent of damage to white matter microstructure due to neuronal loss and can provide complementary information when combined with structural MRI based analyses (Q. Wang et al., 2018a). Structural brain connectome networks are computed from diffusion weighted MRIs where each node is represented as a distinct brain region and the edges are the number of axonal fibers passing between any two regions. They can reveal disruptions in global as well as local connectivity patterns in several disease areas (Bassett & Bullmore, 2009). Graph theory based measures of brain networks have been used to study changes associated with neurological diseases (Bullmore & Sporns, 2009; Rubinov & Sporns, 2010).

Machine Learning (ML) and Deep Learning (DL) methods are extensively used in neuroimaging analyses of Alzheimer's Disease. Classical Machine Learning (ML) approaches involve extracting and selecting features from medical images such as cortical or subcortical volumetric measurements from structural MRIs (sMRI) (Dickerson et al., 2011; Feng & Ding, 2020; Ledig et al., 2018); graph theory features from functional connectivity networks or temporal correlations between brain regions derived from functional MRIs (fMRI) (Damoiseaux, 2012; Sperling, 2011; K. Wang et al., 2007); or abnormal anatomical connectivity patterns from diffusion MRIs (Billeci et al., 2020). With the increase in computational resources in recent years, Deep

Learning based methods, particularly Convolutional Neural Networks have received widespread attention in the medical image analysis community for its ability to automatically extract patterns from raw images. While the predictive performance of Deep Learning in AD diagnosis and prognosis (Ebrahimighahnavieh et al., 2020; Jo et al., 2019; Liu et al., 2014) has been impressive, they are usually black-box models and suffer from lack of interpretability, which is an essential component in biomedical research.

ML/DL based methods generally require a large dataset for training the models and the demand is further amplified in high-dimensional datasets to avoid the curse of dimensionality phenomenon. High dimensional, low sample size and noisy datasets are common in biomedical research. In cases where there are multiple labels, for example NC, EMCI, LMCI, and AD, a multi-class classifier or multiple binary classifiers for each pair of disease stages can be trained depending on the choice of algorithm and availability of data. While multi-class classifiers such as neural networks, random forests, multinomial regression, etc. allow ease during training, they may favor certain classes that are either more representative or have better separability (Sánchez-Maroño et al., 2010). Training multiple binary classifiers instead allows flexibility on the model parameters for each binary classifier, however these binary classifiers are often trained independently of each other. (Li et al., 2014; Shalev-Shwartz et al., 2011; Zweig & Weinshall, 2013) have proposed methods for information sharing and joint learning for classification, however, to our knowledge none have been used in the context of Alzheimer's Disease classification.

In this thesis, we propose a new framework for Shared Simultaneous Learning of sparse logistic regression classifiers for classification of NC, EMCI, LMCI and AD using anatomical whole brain networks extracted from diffusion MRI. We further identify the connectivity patterns

associated with disease progression. Section 2.0 below discusses the necessary background, followed by the description of the Shared Simultaneous Learning algorithm and Alzheimer's Disease Neuroimaging Initiative 2 (ADNI2) dataset used in this thesis in Section 3.0. Experimental setup and results are discussed in Section 4.0 and we conclude with the conclusions and future scope in Section 5.0.

2.0 Background

2.1 Logistic Regression

Logistic Regression is a special case of Generalized Linear Model (GLM), more specifically, it is Binomial Regression with Logit Link (McCullagh & Nelder, 1983; Rodríguez, 2007). GLM models the conditional expected value of the outcome variable Y as a linear combination of the independent variables $[X_1, X_2, \dots, X_p]$,

$$\boldsymbol{E[Y|X] = \beta_0 + \sum_{j=1}^{p} \beta_j X_j} \qquad \textbf{Equation 2-1}$$

where X_j is the j^{th} independent variable, β_0 is the bias term and β_j are coefficients that are to be estimated from the data.

In the simple scenario above, GLM is the familiar Linear Regression where the outcome variable is continuous and can take on any real value. However, when the outcome is categorical, and if it can take only one of two values such as Success/Failure, Yes/No, Pass/Fail, Presence/Absence, etc., the outcome is coded as 1 for the positive or desired class and 0 for the negative class. The outcome variable Y_i for i^{th} observation follows a Bernoulli Distribution with probability of desired class π. Bernoulli Distribution can be written as

$$\boldsymbol{P[Y_i = y_i] = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}} \qquad \textbf{Equation 2-2}$$

If observations with identical values of independent variables (covariates) can be grouped into k groups with m_i number of observations in the i^{th} group, the outcome variable $Y_i$can be modelled as a realization of the Binomial Random Variable such that

$$\boldsymbol{P[Y_i = y_i] = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}} \qquad \textbf{Equation 2-3}$$

The Bernoulli distribution is a special case of Binomial distribution when $m_i = 1$, and in practice, observations with identical independent variables are de-aggregated and both approaches lead to the same likelihood function which will be discussed further in Section 2.1.1 below. Thus, the conditional expected value of the outcome is the conditional probability of desired class in a single Bernoulli trial given the independent variables,

i.e. $$\boldsymbol{E[Y|X] = P[Y_i = 1|X_i] = \pi_i(x) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}} \qquad \textbf{Equation 2-4}$$

The model of the form as described in Equation 2-1 is inappropriate in this case as the outcome can take any real value. The outcome in Equation 2-4 is a probability value of the positive class which is between [0,1]. In order to map the outcome variable to a real value, the logit link is used which is given by

$$\boldsymbol{logit(\pi_i(x)) = ln\left(\frac{\pi_i(x)}{1 - \pi_i(x)}\right)} \qquad \textbf{Equation 2-5}$$

It is obvious to note that as the probability of positive class π_i approaches zero, the logit link described in Equation 2-5 approaches -∞ and vice versa. Hence, we get a mapping of the probability values between [0, 1] to a real value in [-∞, +∞] (Figure 2).

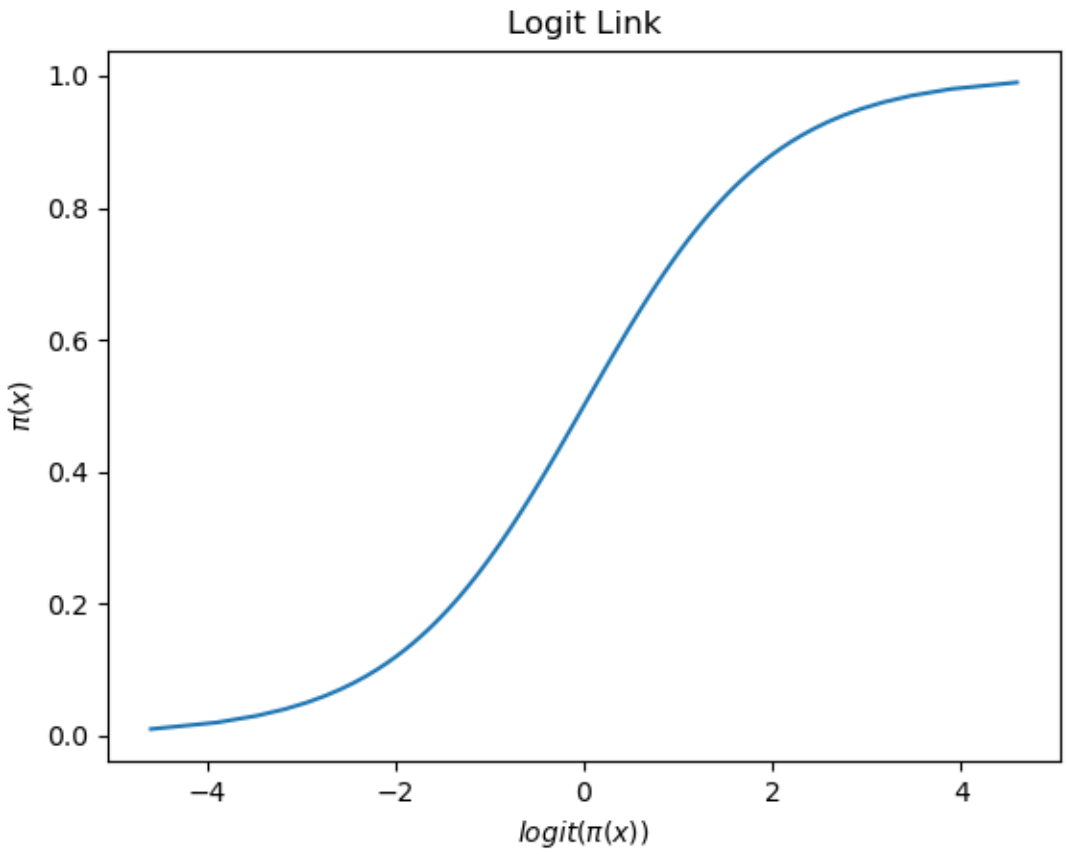

Figure 2: Logit Link mapping probability value between 0 and 1 to real axis

The logistic regression model can thus be written as

$$\boldsymbol{logit}(\boldsymbol{\pi_i}(\boldsymbol{x})) = \boldsymbol{\beta_0} + \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\beta_j X_j} \qquad \textbf{Equation 2-6}$$

We can rewrite Equation 2-6 in terms of positive class probability π_i as

$$\boldsymbol{\pi_i}(\boldsymbol{x}) = \frac{\boldsymbol{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}}{\boldsymbol{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}} = \frac{\boldsymbol{1}}{\boldsymbol{1 + e^{-\beta^T X}}} \qquad \textbf{Equation 2-7}$$

where $\boldsymbol{\beta}^T X$ is the vectorized form of $\boldsymbol{\beta_0} + \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\beta_j X_j}$ after appending a 1 to the observation vector (McCullagh & Nelder, 1983; Rodríguez, 2007).

Logistic Regression has traditionally been used in statistical inference to model the effect of independent variables when the outcome is categorical. It outputs a probability value for the positive class using a function of linear combination of the independent variables as per Equation 2-7. This property allows logistic regression to be used as a classification algorithm when a threshold is set on the estimated class probability. In practice, for balanced data sets this threshold is usually set to 0.5 i.e. if the output probability is greater than 0.5, the observation is assigned to the positive class and vice versa. Thus, Logistic Regression can be used as a linear classifier where the decision boundary separating the two classes is a $d-1$ dimensional hyperplane for d-dimensional data.

2.1.1 Maximum Likelihood Estimation

Logistic Regression estimates the outcome as a probability of desired class given the vector of independent variables. For one observation, we can write this probability as the Bernoulli Distribution as in Equation 2-2. For n independent and identically distributed observations, the likelihood function can be written as

$$\boldsymbol{L(\beta) = \prod_{i=1}^{n} \pi_i(x)^{y_i} \ (1 - \pi_i(x))^{1-y_i}} \qquad \textbf{Equation 2-8}$$

The idea is to maximize this likelihood function with respect to the coefficient vector β. Note that the probability $\pi_i(x)$ is a function of the independent variables and the coefficients as per Equation 2-7. Instead of maximizing Equation 2-8 directly, it is convenient to maximize the natural logarithm of the likelihood function as logarithm is a monotonically increasing function. The log likelihood can be written as

$$\boldsymbol{l^{LR}(\beta) = \sum_{i=1}^{n} y_i \ln(\pi_i(x)) + (1 - y_i)\ln(1 - \pi_i(\mathrm{x}))} \qquad \textbf{Equation 2-9}$$

It is interesting to note that if identical covariates were grouped and the outcome was modelled as a Binomial Distribution, the log-likelihood would be

$$\boldsymbol{l(\beta) = \sum_{i=1}^{n} \binom{m_i}{y_i} + y_i \ln(\pi_i(x)) + (1 - y_i)\ln(1 - \pi_i(\mathrm{x}))} \qquad \textbf{Equation 2-10}$$

The first term in Equation 2-10 does not depend on the coefficient vector, therefore maximizing Equation 2-10 is the same as that of Equation 2-9.

The coefficients enter the likelihood function in a non-linear way through the logistic function. Hence, a closed form solution by setting the first derivative equal to zero does not exist for maximizing the likelihood function. Instead, numerical optimization methods are used for estimating the coefficient vectors. The logistic regression likelihood is a twice differentiable concave function which makes it suitable for Newton's Method for Numerical Optimization. The simplest and most common method is called the Gradient Descent which is described below.

2.1.2 Gradient Descent

Gradient Descent is a simple yet powerful optimization algorithm for finding the minimum of smooth objective functions. The main idea is to take repeated steps in the direction of negative gradient at the current point. Since the logistic log-likelihood is smooth and concave, gradient descent is well suited to minimize the negative of the log-likelihood function. If and when the algorithm reaches the global minimum, the gradients, in theory, are zero and the algorithm is said to have converged. In practice, the algorithm is said to have converged if the gradients are less than a small value ε. The vector parameter update rule is given by

$$\boldsymbol{\beta^{t+1} = \beta^{t} - \eta \nabla l(\beta^{t})} \qquad \textbf{Equation 2-11}$$

where $\beta^t = \left[\beta_0^t, \beta_1^t, \dots, \beta_p^t\right]^T$ is the $(p+1) \times 1$ dimensional parameter vector at step t, $\eta > 0$ is the step size also known as the learning rate, $\nabla l^{LR}(\beta) = \left[\frac{\partial l^{LR}(\beta^t)}{\partial \beta_0^t}, \frac{\partial l^{LR}(\beta^t)}{\partial \beta_1^t}, \dots, \frac{\partial l^{LR}(\beta^t)}{\beta_p^t}\right]^T$ is the vector of gradients or partial derivatives of the logistic log-likelihood (Equation 2-9) with respect to each parameter β_j, and $l^{LR}(\beta^t)$ is the log-likelihood (Equation 2-9) using parameters obtained at step t.

The gradient vector for the logistic log-likelihood is computed using

$$\boldsymbol{\nabla l^{LR}(\beta) = \sum_{i=0}^{n} [y_i - \pi_i(x)]\, x_i} \qquad \textbf{Equation 2-12}$$

y_i is the true class label, $\pi_i(x)$ is the predicted class probability obtained from Equation 2-7, and $x_i = \left[1, x_{i1}, x_{i2}, \dots, x_{ip}\right]$ is the i^{th} observation vector with a 1 as the first element for the bias term.

2.1.3 Interpreting Logistic Regression Coefficients - Odds Ratio

For a simple linear regression, let β be the coefficient associated with the independent variable. Therefore, a one unit change in the independent variable will lead to an average change of β units in the outcome variable. The interpretation of coefficients in logistic regression is not as straightforward as that in case of linear regression due to the logit link Equation 2-5 being modelled as a linear combination of the independent variables instead of the class probabilities. The logit link is the natural logarithm of the ratio of the probability of positive class $\pi_i(x)$ to the probability of negative class $1 - \pi_i(x)$. This ratio is called as "odds" and logistic regression is linear in terms

of log odds. An intuitive example for odds is if the probability of a biased coin landing heads is 0.6, the odds of landing a heads are 1.5 times that of landing a tails.

From Equation 2-5 and Equation 2-6, the logistic regression model is specified by

$$\boldsymbol{ln\left(\frac{\pi_i(x)}{1-\pi_i(x)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j} \qquad \textbf{Equation 2-13}$$

Exponentiating, we get the odds as

$$\boldsymbol{\frac{\pi_i(x)}{1-\pi_i(x)} = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_j\right)} \qquad \textbf{Equation 2-14}$$

To see the effect of 1-unit change in one independent variable say X_1 assuming all other variables are held constant, we can compute the odds ratio (OR) as

$$\mathbf{OR} = \frac{\exp\left(\beta_0 + \beta_1(X_1 + 1) + \beta_2 X_2 + \cdots + \beta_p X_p\right)}{\exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right)}$$

$$\mathbf{OR} = \exp(\beta_1) \qquad \textbf{Equation 2-15}$$

Therefore, assuming all other variables are held constant, a 1 unit increase in one variable say X_1 leads to a β_1 unit increase in the log-odds or equivalently $\exp(\beta_1)$ units increase in the odds (Molnar, 2019). If the coefficient β_1 is negative, $\exp(\beta_1)$ will be less than 1. In this case a 1 unit increase in the variable, of course assuming all other variables are held constant, will "decrease" the odds by a factor of $\exp(\beta_1)$. This makes Logistic Regression a powerful and interpretable machine learning algorithm.

2.1.4 Regularization in Logistic Regression

In many real-world scenarios, sample size may be limited, or the number of independent variables could be bigger than available data. Many variables may also have high correlation, which introduces the problem of multicollinearity. Further, if any independent variable perfectly separates the classes, the maximum likelihood estimate does not exist and the optimization fails to converge. Also, logistic regression is prone to overfitting in that there is a perfect fit on the training set yet poor generalization on the test set, especially in high dimensions. These problems can be mitigated by introducing regularization or penalty on the coefficients. Regularization is usually achieved by adding a penalty term $r(\beta)$ to the negative log likelihood function

$$\boldsymbol{l(\beta) = -\sum_{i=1}^{n} y_i \ln(\pi_i(x)) + (1 - y_i) \ln\left(1 - \pi_i(\mathbf{x})\right) + \lambda\, r(\beta)} \quad \textbf{Equation 2-16}$$

where λ is a tuning parameter that controls the amount of regularization. In practice, λ is selected by performing cross validation on the training set.

2.1.4.1 L2 Regularization

L2 penalty is also called as Ridge regularization where the $\ell 2$ or squared Euclidean norm of the coefficient vector $||\beta||_2^2$ is used as the penalty term (Hoerl & Kennard, 2000). The loss function thus becomes

$$\boldsymbol{l(\beta) = -\sum_{i=1}^{n} y_i \ln(\pi_i(x)) + (1 - y_i) \ln\left(1 - \pi_i(\mathbf{x})\right) + \lambda\, ||\beta||_2^2} \quad \textbf{Equation 2-17}$$

The L2 regularization assigns a small non-zero value to the coefficients to achieve a smaller value of the loss. The L2 regularization shrinks the coefficients towards zero and is therefore also referred to as shrinkage estimator. Increasing the value of λ will result in more coefficients being

pushed towards zero, however none of the coefficients will achieve a value of exactly zero. Because the squared Euclidean norm is both smooth, differentiable, and convex, Newton methods such as Gradient Descent described above can be used for optimization. In the Bayesian framework, L2 regularization can be interpreted as having a Normally distributed prior on the coefficients with zero mean and variance $\frac{1}{2\lambda}$.

2.1.4.2 L1 Regularization

In many cases such as when the number of independent variables is far greater than the number of available data, a sparse solution is desired where most coefficients are assigned a value of zero. The Least Absolute Shrinkage and Selection Operator (LASSO) regularization (Tibshirani, 1996), also known as L1 regularization is used for automated variable selection and regularization where the $\ell 1$ norm of the coefficient vector, $||\beta||_1$, is used as the penalty term. The likelihood/loss function thus becomes

$$\boldsymbol{l(\beta) = -\sum_{i=1}^{n} y_i \ln\left(\pi_i(x)\right) + (1 - y_i)\ln\left(1 - \pi_i(\mathrm{x})\right) + \lambda\, ||\beta||_1} \quad \textbf{Equation 2-18}$$

The L1 regularization results in a sparse estimate where some of the coefficients are assigned a value of exactly zero. As with L2 regularization, increasing the value of λ will result in more variables being discarded. When two variables are correlated, LASSO picks one of them at random even though both might have predictive power, whereas the L2 regularization shrinks both towards each other. Elastic net (Zou & Hastie, 2005) overcomes this by combining both the L1 and L2 penalties. The L1 norm is convex but not differentiable at 0, therefore, Newton methods such as Gradient Descent described above cannot be used for optimization. Coordinate Descent is one of

the popular algorithms used for solving LASSO Logistic Regression (Friedman et al., 2010). In the Bayesian framework, L1 regularization can be interpreted as having a Laplacian prior on the coefficients with zero mean and scale parameter $\frac{1}{\lambda}$.

2.2 Model Evaluation Metrics

Logistic Regression outputs a probability value between 0 and 1 for the positive class. After setting a threshold on this probability, logistic regression can be used as a classification algorithm. Generally, for balanced datasets this threshold is set to 0.5, thus observations with probability greater than 0.5 are classified into Class 1 and those below 0.5 are classified into Class 2. The accuracy can then be simply computed by comparing the number of correct predictions with the ground truth labels. However, say for example logistic regression predicts a probability of 0.49 for an observation from Class 0. The 0.5 threshold in this case will correctly assign the observation to Class 0 but the accuracy will be an overoptimistic measure of the model performance as random factors could influence the prediction probability. Further, for imbalanced datasets where majority of observations are from a particular class, if the model predicts all observations as the majority class, the accuracy will be biased estimate of model performance.

Area Under the Receiver Operating Characteristic Curve (AUROC) or simply Area Under the Curve (AUC) is an unbiased measure that is used to evaluate the confidence of model predictions irrespective of class distribution. First, a prediction is True Positive (TP) when, given that the observation is from the positive class, the model prediction is also positive whereas if the model prediction is negative, it is termed as False Negative (FN). Similarly, if the observation is from the negative class and the model prediction is also negative, it is termed as True Negative

(TN). If the model prediction is positive for an observation from the negative class, it is counted as False Positive (FP)(Fawcett, 2006). Table 1 summarizes the concept of TP, TN, FP and FN using a confusion matrix.

Table 1: Confusion Matrix

True Label \ **Predicted Label**	***Positive***	***Negative***
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

From the confusion matrix, several metrics can be defined of which the Accuracy, Sensitivity and Specificity are given below.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \qquad \textbf{Equation 2-19}$$

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}} \qquad \textbf{Equation 2-20}$$

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN + FP}} \qquad \textbf{Equation 2-21}$$

Sensitivity is the ratio of number of True Positives to the number of all positive class observations and Specificity is the ratio of number of True Negatives to the number of all negative class observations. The ROC curve is the plot between Sensitivity versus (1-Specificity) as the threshold on the class probability is varied from 0 to 1. AUROC can provide a point estimate to summarize the ROC curve and it range is [0,1] with an AUC of 1 depicting an ideal classifier with highest confidence in its predictions. An AUC of 0.5 corresponds random predictions from the classifier.

While it is ideally desired that the AUC value be between 0.5 and 1, the low sample size, low effect size data which are typical in neuroscience studies can result in below chance level performance of classifiers (Jamalabadi et al., 2016). Figure 3 shows the ROC with different values of AUC.

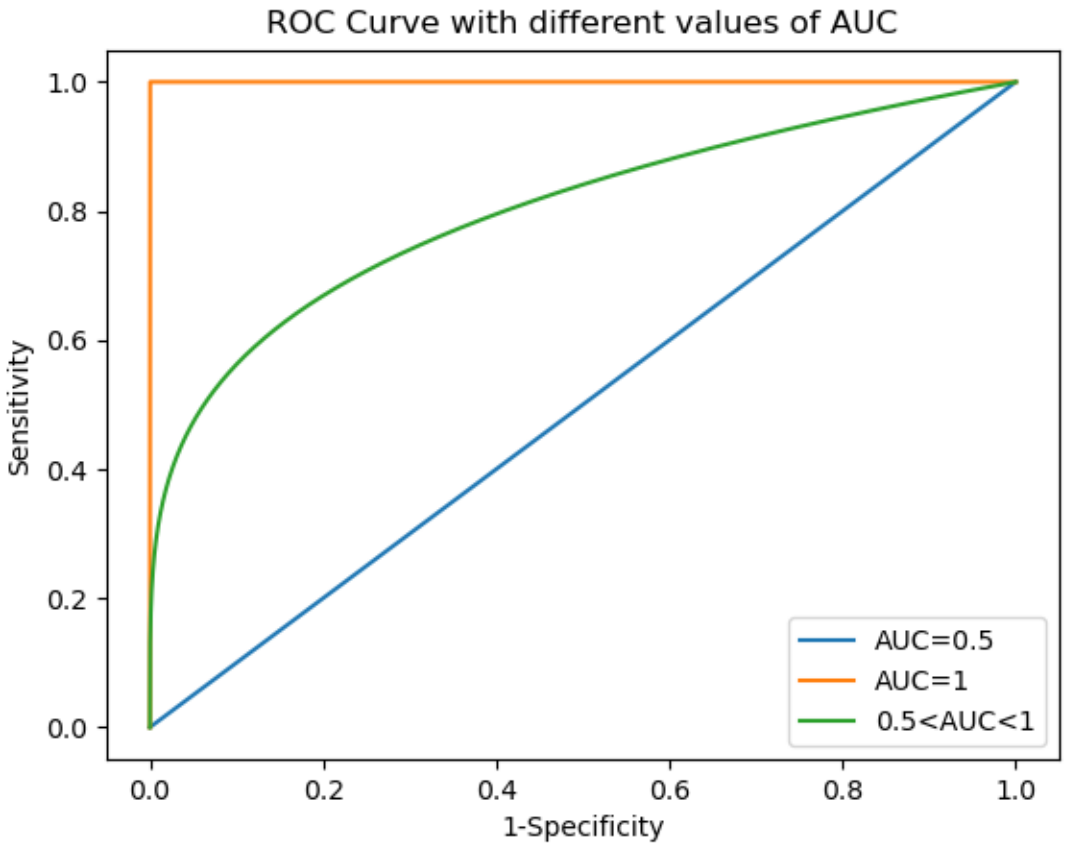

Figure 3: ROC Curve with Different Values of AUC

3.0 Methods

3.1 Dataset

Diffusion Weighted Magnetic Resonance Imaging (dMRI) and T1 Weighted MRI scans used to extract the structural connectomes used in the thesis were obtained from the second stage of Alzheimer's Disease Neuroimaging Initiative (ADNI2) database (adni.loni.usc.edu). ADNI is a multi-institutional longitudinal study launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. Over the years, the data generated from ADNI has resulted in the extensive research of biomarkers from serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment of a combination of them to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Diffusion Weighted and T1 MRI scans were collected from 202 participants across 16 sites in the United States and Canada. The detailed inclusion and exclusion criteria is available in the ADNI2 protocol which can be accessed at http://adni.loni.usc.edu/wp-content/uploads/2008/07/adni2-procedures-manual.pdf as of June 29, 2021. Subjects were scanned on a 3T General Electric Medical Systems scanner with the following parameters

- **3D T1**-weighted images generated using spoiled gradient echo (SPGR) sequences
    - **Matrix Size:** 256 x 256
    - **Voxel Size:** 1.2 x 1.0 x 1.0 mm^3
    - **Inversion Time (TI):** 400 ms

- **Repetition Time (TR):** 6.98 ms
- **Echo Time (TE):** 2.85 ms
- **Flip Angle:** 11º

- **Diffusion Weighted Imaging (DWI) Images**
  - **Matrix Size:** 128 x 128
  - **Voxel Size:** 2.7 x 2.7 x 2.7 mm^3
  - **Repetition Time (TR):** 9050 ms
  - **Number of Slices:** 59
  - **Scan Time:** 9 minutes

For each DWI image, 46 scans were acquired per subject consisting of 5 T2 weighed images with no diffusion sensitization (b_0 images) and 41 DWI scans with b=1000 s/mm^2. Additional image acquisition details can be found in the ADNI2 MRI Protocols at http://adni.loni.usc.edu/wp-content/uploads/2010/05/ADNI2_GE_3T_22.0_T2.pdf as of June 29, 2021. All scans were included after performing quality assurance through visual assessment of both T1 and DWI images.

The dataset consisted of 51 Normal Controls (NC), 73 Early Mild Cognitive Impairment (EMCI), 39 Late Mild Cognitive Impairment (LMCI), and 39 Alzheimer's Disease (AD) subjects. The diagnostic criteria were based on the Mini-Mental State Exam (MMSE), Clinical Dementia Rating (CDR), and education adjusted scores from Logical Memory II subscale from the Wechsler Memory Scale (LM2-WMS). Subject demographics and diagnostic criteria are outlined in Table 2 and Table 3 respectively.

Table 2: Subject Demographics

Disease Stage	Number	Age (in years)	Sex
NC	51	72.42(±6.15)	M: 22 F: 29
EMCI	73	72.43(±8.00)	M: 47 F: 26
LMCI	39	72.32(±5.82)	M: 24 F: 15
AD	39	75.56(±9.11)	M: 25 F: 14
Total	202	73.01(±7.48)	M: 118 F: 84

Table 3: Diagnostic Criteria

Disease Stage	MMSE	CDR	LM2-WMS	Other Criteria
NC	24 to 30	0	≥9 for ≥16 of education ≥5 for 8-15 years of education ≥3 for 0-7 years of education	Cognitively normal No memory complaints
EMCI	24 to 30	0.5	9-11 for ≥16 of education 5-9 for 8-15 years of education 3-6 for 0-7 years of education	No significant cognitive impairment Subjective memory concern
LMCI	24 to 30	0.5	≤8 for ≥16 of education ≤4 for 8-15 years of education ≤2 for 0-7 years of education	No significant cognitive impairment Subjective memory concern
AD	20 to 26	0.5 or 1.0	≤8 for ≥16 of education ≤4 for 8-15 years of education ≤2 for 0-7 years of education	NINCDS/ADRDA criteria for probable AD

3.2 Data Pre-processing

3.2.1 Preprocessing Raw MRI Data

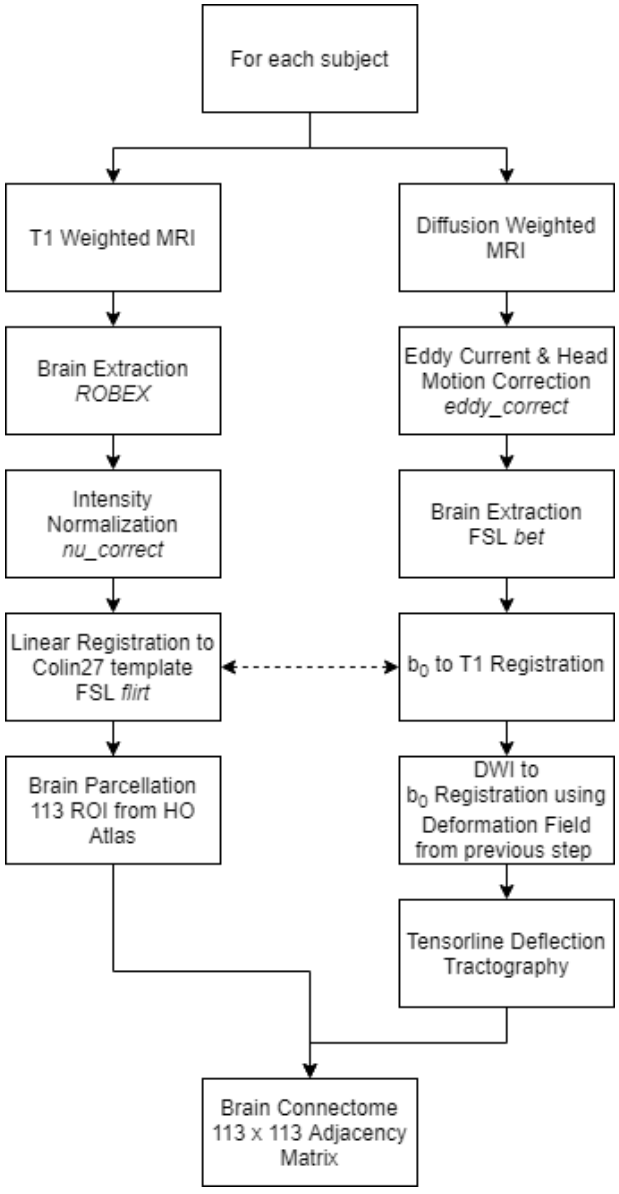

Figure 4: Raw MRI Data Preprocessing Pipeline

After acquiring the T1 weighted and Diffusion Weighted MRI scans, preprocessing was performed on data from each subject. Figure 4 shows the flowchart of preprocessing performed originally in Zhan et al., 2015. First, skull stripping was performed on T1 MRI using ROBEX (Iglesias et al., 2011) to extract brain tissue followed by visual inspection for manual edits if needed. Intensity normalization was performed using the MNI *nu_correct* (*The Brain Imaging Software Toolbox*, n.d.) tool to correct for intensity fluctuations. T1 MRI from all subjects were aligned to the same 3D space by registering them to Colin27 Brain Template (Holmes et al., 1998) using FSL *flirt* (Jenkinson et al., 2002; Jenkinson & Smith, 2001).

For the DWI images, the FSL *eddy_correct* (https://fsl.fmrib.ox.ac.uk/fsl) (Smith et al., 2004) tool was used to correct for head motion and eddy current distortions. This preprocessing was performed before a more robust tool for eddy current correction, FSL *eddy* (Andersson & Sotiropoulos, 2016) was available. Skull stripping was then performed using FSL Brain Extraction Tool (BET) to remove non-brain tissue (Smith, 2002). Further, echo-planar induced (EPI) susceptibility artifacts were corrected by linearly aligning and then elastically registering the b_0 images to the pre-processed T1 scans using an inverse consistent registration algorithm with a mutual information cost function (Leow et al., 2007). The final step of preprocessing raw DWI images consisted of applying the 3D deformation fields obtained from previous step to the 41 DWI scans.

3.2.2 Computing Brain Connectome Networks

The brain connectome network for each subject were computed using the pre-processed T1 and DWI MRI. Each subject's scan was parcellated into 113 cortical and subcortical region of interests (ROIs) based on the Harvard Oxford (HO) Cortical and Subcortical probabilistic atlas

(Desikan et al., 2006). These 113 brain ROIs are listed in Table 4. Left and right hemispheric ROIs for each cortical region were defined by bisecting the midline cortical masks into left and right components, followed by thresholding these masks at 10% in order to include tissue along the Gray Matter-White Matter interface. As the HO atlas is defined for the MNI152 T1 brain template, the affine transformation matrix between MNI152 template (Fonov et al., 2009, 2011) and each subject's skull stripped T1 image as well as between each subject's skull-stripped T1 image and Fractional Anisotropy (FA) image was determined using FSL *flirt* (Jenkinson et al., 2002; Jenkinson & Smith, 2001). Finally, the 113 ROIs from HO atlas were transformed to each subject's DWI space using nearest-neighbor interpolation after combining the two affine transformation matrices obtained in the previous step, and each voxel was assigned to the ROI for which it had the highest probability of membership. Tensorline Deflection (Lazar et al., 2003), a tensor based deterministic tractography method was used to track white matter fibers in the DWI images (R. Wang et al., 2007). The idea is to "deflect" the incoming propagation direction vector of white matter tracts towards the direction of the major eigenvalue using the entire diffusion tensor. The gray matter and cerebrospinal fluid regions were excluded by restricting the tractography to voxels with Fractional Anisotropy (FA) value greater than or equal to 0.2. Further, false positive tracts were avoided by stopping the tractography for pathways with sharp turns of over 30° (Zhan et al., 2013, 2015).

The adjacency matrix for each subject was thus computed with nodes representing 113 brain ROIs as per the HO atlas. The edges between two nodes were computed by counting the fibers identified using the Tensorline Deflection tractography method described above that intersected both ROIs. The brain connectome network thus generated is a simple, undirected graph.

Table 4: List of 113 Brain Region of Interest from Harvard Oxford Atlas

For each entry, the smaller number on the left corresponds to the ROI in left hemisphere and the larger number on the right corresponds to the ROI in right hemisphere of the brain.

Node	Brain ROI	Node	Brain ROI
1, 9	Thalamus	2, 10	Caudate
3, 11	Putamen	4, 12	Pallidum
5	Brainstem	6, 13	Hippocampus
7, 14	Amygdala	8, 15	Accumbens
16, 17	Frontal Pole	18, 19	Insular cortex
20, 21	Superior Frontal Gyrus	22, 23	Middle frontal gyrus
24, 25	Inferior frontal gyrus, pars triangularis	26, 27	Inferior frontal gyrus, pars opercularis
28, 29	Precentral gyrus	30, 31	Temporal pole
32, 33	Superior temporal gyrus, anterior division	34, 35	Superior temporal gyrus, posterior division
36, 37	Middle temporal gyrus, anterior division	38, 39	Middle temporal gyrus, posterior division
40, 41	Middle temporal gyrus, temporooccipital part	42, 43	Inferior temporal gyrus, anterior division
44, 45	Inferior temporal gyrus, posterior division	46, 47	Inferior temporal gyrus, temporooccipital part
48, 49	Postcentral gyrus	50, 51	Superior parietal lobule

Table 4 (continued)

52, 53	Supramarginal gyrus, anterior division	54, 55	Supramarginal gyrus, posterior division
56, 57	Angular gyrus	58, 59	Lateral occipital cortex, superior division
60, 61	Lateral occipital cortex, inferior division	62, 63	Intracalcarine cortex
64, 65	Frontal medial cortex	66, 67	Juxtapositional lobule cortex
68, 69	Subcallosal cortex	70, 71	Paracingulate gyrus
72, 73	Cingulate gyrus, anterior division	74, 75	Cingulate gyrus, posterior division
76, 77	Precuneus cortex	78, 79	Cuneal cortex
80, 81	Frontal orbital cortex	82, 83	Parahippocampal gyrus, anterior division
84, 85	Parahippocampal gyrus, posterior division	86, 87	Lingual gyrus
88, 89	Temporal fusiform cortex, anterior division	90, 91	Temporal fusiform cortex, posterior division
92, 93	Temporal occipital fusiform cortex	94, 95	Occipital fusiform cortex
96, 97	Frontal opercular cortex	98, 99	Central opercular cortex
100, 101	Parietal opercular cortex	102, 103	Planum polare
104, 105	Heschl's gyrus	106, 107	Planum temporale
108, 109	Supracalcarine cortex	110, 111	Occipital pole
112, 113	Cerebellum		

3.3 Shared Simultaneous Learning Framework

In many classification tasks, especially in healthcare, the sample size is severely limited and the dataset may not have enough power to train a multi-class classifier directly. Hence, multiple binary classification schemes such as one vs one or one vs rest are preferred. Whereas these binary classifiers are traditionally trained independently of each other, this thesis proposes a framework that would allow sharing information across the multiple binary classifiers during their training. The assumption of the proposed method is that certain features are consistent in their predictive power across classes and we formulate a framework to allow sharing this information during the training.

Let $X^k \in \mathbb{R}^{n \times (p+1)}$ be the training matrix with n observations and p features along with a column of ones for the intercept term, $Y^k \in \{0,1\}$ be the vector of size $n \times 1$ with class labels and $\beta^k \in \mathbb{R}^{(p+1) \times 1}$ be the coefficient vector for k^{th} classifier, $k \in \{1,2,..,K\}$. The coefficient vector is initialized with zeros. In this thesis, K=3 as three binary classifiers were trained for NC vs EMCI, EMCI vs LMCI and LMCI vs AD classification. A new term is added to the sparse logistic regression likelihood function Equation 2-18 to minimize the squared Euclidean distance between the coefficient vectors of K classes. The loss function to optimize for the k^{th} classifier thus becomes

$$\begin{aligned} \boldsymbol{l^{SSL}(\beta^k)} &= -\boldsymbol{l^{LR}(\beta^k)} + \boldsymbol{\lambda_1}\,||\boldsymbol{\beta^k}||_1 + \sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{\lambda_{k,j}} * ||\boldsymbol{\beta^k} - \boldsymbol{\beta^j}||_2^2 \\ &= -\boldsymbol{l^{LR}(\beta^k)} + \boldsymbol{\lambda_1}\,||\boldsymbol{\beta^k}||_1 \\ &\quad + \sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{\lambda_{k,j}}(\boldsymbol{\beta^k} - \boldsymbol{\beta^j})^{\boldsymbol{T}}(\boldsymbol{\beta^k} - \boldsymbol{\beta^j}) \end{aligned} \qquad \textbf{Equation 3-1}$$

where $l^{LR}(\beta^k)$ is the log likelihood of logistic regression given by Equation 2-9, λ_1 controls the sparsity and $\lambda_{\mathrm{k,j}}$ controls the sharing of weights between the j^{th} and k^{th} classifiers. λ_1 and $\lambda_{\mathrm{k,j}}$ should be selected by performing cross validation on the training set. The intercept term is not included in the LASSO as well as the shared weights penalty term.

The Euclidean distance function is smooth and convex, however the $\ell 1$ norm is not differentiable at zero. In order to use gradient based methods for minimizing the proposed loss in Equation 3-1, the $\ell 1$ norm is approximated with the following smooth function (Lee et al., 2006)

$$||\boldsymbol{\beta}||_{\mathbf{1}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} |\boldsymbol{\beta}|$$

$$\approx \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \sqrt{\boldsymbol{\beta_j^2 + \mathcal{E}}} \qquad \textbf{Equation 3-2}$$

for a small value of $\mathcal{E}$. In this work, we set $\mathcal{E} = 1\mathrm{e}^{-10}$. It must be noted that the $\ell 1$ norm is only computed for the coefficients associated with the independent variables, therefore the summation goes from "1" to "p" i.e. the intercept term β_0 is not included in the penalty.

Let $\nabla L1$ be the gradient vector of the smooth approximated $\ell 1$ norm given by Equation 3-2. Thus, the j^{th} element of this vector i.e. the partial derivative of the smooth approximated $\ell 1$ norm with respect to the coefficient β_j will be

$$\boldsymbol{\nabla L1_j = \frac{\partial ||\beta||_1}{\partial \beta_j} = \frac{\beta_j}{\sqrt{\beta_j^2 + \mathcal{E}}}} \qquad \textbf{Equation 3-3}$$

To compute the gradient vector $\nabla l^{SSL}(\beta^{k,t})$ for the k^{th} classifier at time step t for the proposed loss defined in Equation 3-1, the coefficient vectors of the remaining $K-1$ classifiers obtained at the most recent time step are shared. The gradient computation is thus

$$\nabla \boldsymbol{l}^{*}\left(\boldsymbol{\beta}^{\boldsymbol{k,t}}\right) = -\nabla \boldsymbol{l}^{\boldsymbol{LR}}\left(\boldsymbol{\beta}^{\boldsymbol{k,t}}\right) + \boldsymbol{\lambda_1}\nabla \boldsymbol{L1} + \mathbf{2}\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{\lambda}_{\mathbf{k,j}}\left(\boldsymbol{\beta}^{\boldsymbol{k,t}} - \boldsymbol{\beta}^{\boldsymbol{j,t}}\right) \qquad \textbf{Equation 3-4}$$

where $\nabla l^{LR}(\beta^{k,t})$ is the gradient of the logistic log likelihood defined by Equation 2-12 and $\nabla L1$ is the gradient of the smooth approximated $\ell 1$ penalty as defined in Equation 3-3.

Gradient Descent is used to minimize this loss function for each classifier. The gradient descent update rule for the k^{th} classifier can be written from Equation 2-11 as

$$\boldsymbol{\beta}^{\boldsymbol{k,t+1}} = \boldsymbol{\beta}^{\boldsymbol{k,t}} - \boldsymbol{\eta}\nabla \boldsymbol{l}^{*}(\boldsymbol{\beta}^{\boldsymbol{k,t}}) \qquad \textbf{Equation 3-5}$$

The learning rate η is selected using line search. Table 5 summarizes the proposed Shared Simultaneous Learning (SSL) Algorithm. First, the gradient vectors $\nabla l^{SSL}(\beta^{k,t})$ are computed for all K classifiers. During the computation of the gradient vector for the k^{th} classifier, the coefficient vectors $\beta^{j,t}, j \neq k$ from the remaining $K-1$ are shared to compute the gradient of the proposed Euclidean Distance penalty term. Then the convergence is checked by checking if $\ell 2$ norm of the gradient vectors of all K classifiers is less than a small value ε. If the algorithm has not converged, the coefficient vectors for all K classifiers are updated using the update rule in Equation 3-5. The K coefficient vectors are thus optimized simultaneously.

For training two binary classifiers simultaneously ($K = 2$), the corresponding coefficient vectors are $\beta^1, \beta^2 \in \mathbb{R}^{(p+1)\times 1}$. Consider the coefficient β_j^k associated with the independent variable X_j. If X_j is highly predictive of the class label for both the classifiers, vanilla logistic regression will assign a higher weight to both the parameters β_j^1 and β_j^2. Conversely, a lower weight will be assigned or even discarded by the LASSO penalty if it is not predictive in both classifications. In the above two scenarios, the squared distance between the coefficients β_j^1 and β_j^2 will be low. Thus, adding the proposed Euclidean distance penalty term to the logistic loss

function will have a small impact on the estimate of coefficient β_j^k for both classifiers. Now, if X_j is found to be highly predictive in only one classifier, this information will be shared with the other classifier under the assumption stated above that it may also hold some predictive power for the other classifier. Thus, the squared difference between β_j^1 and β_j^2 will be high and they will be pushed towards each other by the proposed penalty term. For an optimal value of the information sharing parameter $\lambda_{\mathrm{k,j}}$ selected using cross-validation, we hypothesize improvement in the predictive performance. This idea can be generalized to multiple classifiers. The proposed approach of information sharing and simultaneous optimization of multiple binary classifiers thus combines the advantages of multi-class classifiers, which inherently share information by modelling multi-class classification as a single optimization problem, and that of the multiple binary classifiers approach, which allows flexibility on the parameters for individual binary classifiers.

Table 5: Shared Simultaneous Learning Algorithm

Algorithm 1: Shared Simultaneous Learning

(1) **Initialize:** Coefficient vectors $\beta^1, \beta^2, \dots, \beta^K = \vec{0}_{(p+1)\times 1}$

(2) **Input:** Data $X^1, X^2, \dots, X^K,\ X^k \in \mathbb{R}^{n\times(p+1)}$ | Labels $Y^1, Y^2, \dots, Y^K,\ Y^k \in \{0,1\}^{n\times 1}$

(3) **for** (t = 0 to MaxIterations)

(4) **for** (k = 1 to K)

(5) Compute gradients $\boldsymbol{\nabla l^{SSL}}\left(\boldsymbol{\beta^{k,t}}\right)$ for k^{th} classifier using Equation 3-4

(6) **if** $\ell 2\_Norm\left(\boldsymbol{\nabla l^{SSL}}\left(\boldsymbol{\beta^{k,t}}\right)\right) \leq \varepsilon$ for all k

(7) break;

(8) **end if**

(9) **end for**

(10) **for** (k = 1 to K)

(11) Update parameters $\boldsymbol{\beta^{k,(t+1)}}$ using Equation 3-5. Select η using Line Search that minimizes the loss given by Equation 3-1

(12) **end for**

(13) **end for**

4.0 Experimental Results

Distinguishing between adjacent disease stages such as NC vs EMCI or EMCI vs LMCI is challenging as the brain connectivity changes are subtle and detecting the transition from one disease stage to the next is critical for clinicians to take appropriate decisions with regards to patient management and treatment regime. Therefore, this work evaluates 3 classifiers for the adjacent disease stages, NC vs EMCI, EMCI vs LMCI, and LMCI vs AD using the anatomical brain connectome networks derived from Diffusion Weighted Images acquired through ADNI2. The edges from the adjacency matrices which represent unique brain connections were pre-processed, vectorized, and used as features for the SSL algorithm. Section 4.1 below explains the additional pre-processing performed on the adjacency matrices and experimental results on whole brain network are discussed in Section 4.2.

4.1 Preprocessing the Adjacency Matrices

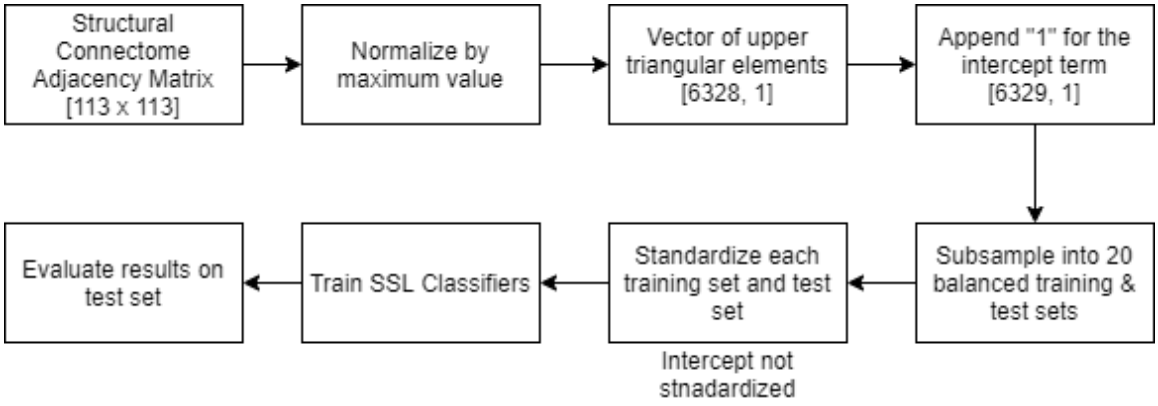

Figure 5: Final Pre-processing Steps for Generating the Training and Test Datasets

Additional pre-processing was performed on the structural brain connectome networks before training the classifiers. Figure 5 shows the final pre-processing steps on the adjacency matrices obtained from the methods described in Section 3.2. Since the scale and range of the networks for each subject is different, each network was normalized by dividing with the maximum value in the adjacency matrix. The first row in Figure 6 shows the normalized adjacency matrices averaged across all subjects in each disease stage NC, EMCI, LMCI, and AD. The second row shows the difference in connectivity patterns across NC vs EMCI, EMCI vs LMCI, and LMCI vs AD. From row 2, the difference between NC-EMCI appears very subtle whereas it is more prominent for LMCI-AD.

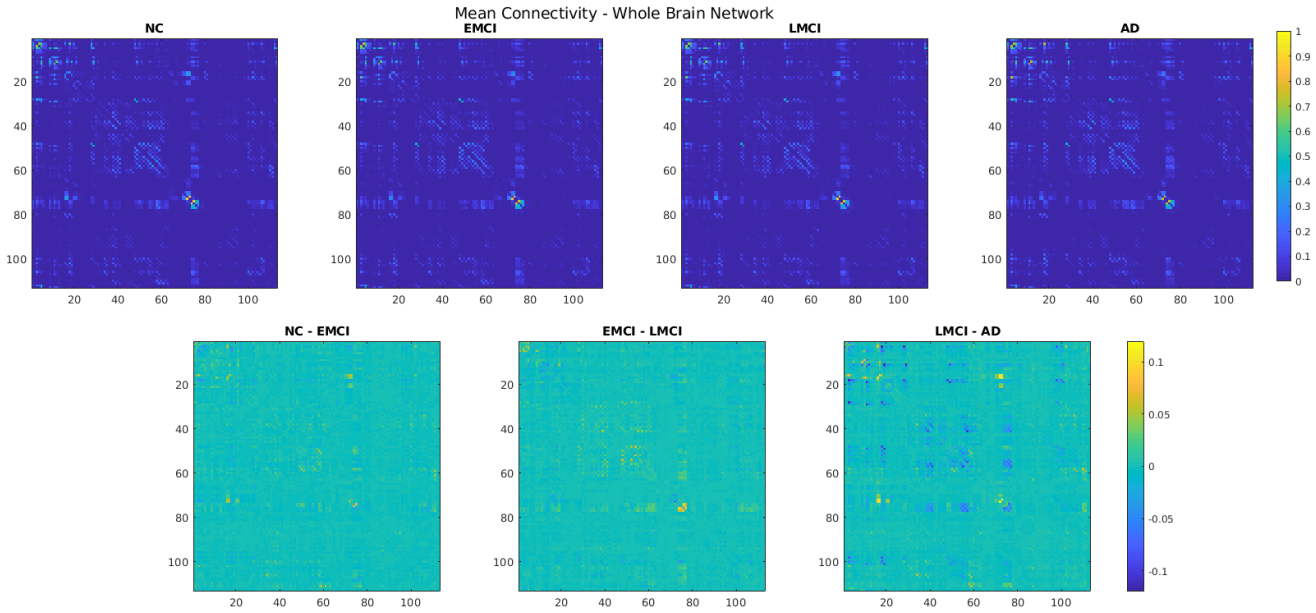

Figure 6: Whole Brain Network. Row 1 - Mean Adjacency Matrices for NC, EMCI, LMCI, and AD. Row 2 – Difference between Mean Adjancency Matrices for (NC - EMCI), (EMCI - LMCI), and (LMCI - AD) Each index on the X-axis and Y-axis corresponds to one of the 113 brain ROIs.

For each subject, the adjacency matrix $A \in \mathbb{R}^{113 \times 113}$ is symmetric and has a zero-diagonal due to no self-connections. Therefore, the upper triangular elements excluding the diagonal, which represent unique connections between the brain regions are extracted and vectorized to form the observation vector $X_i \in \mathbb{R}^{6328 \times 1}$ for the i^{th} subject. A "1" is appended to each observation for the intercept term which yields the final observation vector $X_i \in \mathbb{R}^{6329 \times 1}$ for the i^{th} subject. To avoid bias due to data imbalance, 20 training and test sets are sampled from the entire data such that each class has equal number of observations. The train test split is performed with 85% data in the training set and 15% in the test set. For example, between EMCI and LMCI, 39 observations are randomly sampled from each class of which 33 observations are randomly assigned to the training set and the remaining 6 observations to the test set.

The number of independent variables or features, *p*, are greater than the number of observations, hence the L1 regularization is used to avoid overfitting and generate a sparse model to pick only the relevant features. Each feature in the training data is standardized by subtracting the mean and dividing by the variance to ensure all terms are penalized by the regularization equally. The mean and variance from the training set are used to standardize the cross-validation set as well as the test set to avoid bias. Finally, the Shared Simultaneous Learning Classifiers were trained using best hyperparameters obtained from grid search cross validation and the results are evaluated using accuracy and AUC obtained on the test sets.

4.2 SSL on Structural Brain Connectome

Three classifiers for NC vs EMCI, EMCI vs LMCI and LMCI vs AD were trained using Shared Simultaneous Learning (SSL) Algorithm on the preprocessed data. The upper triangular elements of the 113 × 113 adjacency matrix for each subject were vectorized and used as input for the classifiers. These 113 nodes correspond to 56 ROIs each in the left and right hemisphere of the brain and one ROI for brainstem. Thus, the whole brain connectome was used for training. In this thesis, the information sharing parameter $\lambda_{k,j}$ was set to a constant value λ_2 i.e. $\lambda_{k,j} = \lambda_2$. The optimal hyperparameters λ_1 and λ_2 were identified using 5-fold cross validation repeated 5 times on one training set. For each hyperparameter, grid search was performed using values on the log scale between 10^{-5} to 10. Model performance was then evaluated using mean accuracy and mean AUC on the 20 balanced test sets. The performance of SSL was compared with baseline methods; L1 regularized logistic regression and linear Support Vector Machines (SVM). Table 6 summarizes the results.

Table 6: Classification Results for Whole Brain Connectome

	SSL		L1 Logistic Regression		Linear SVM	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
NC vs EMCI	**0.53 ± 0.13**	**0.53 ± 0.13**	0.51 ± 0.15	**0.53 ± 0.15**	0.47 ± 0.04	0.47 ± 0.11
EMCI vs LMCI	**0.66 ± 0.09**	**0.68 ± 0.11**	0.61 ± 0.15	0.62 ± 0.15	0.52 ± 0.21	0.48 ± 0.23
LMCI vs AD	**0.65 ± 0.11**	**0.73 ± 0.12**	0.61 ± 0.10	0.66 ± 0.11	0.55 ± 0.07	0.62 ± 0.10

SSL outperformed the L1 Logistic Regression and Linear SVM in terms of accuracy and AUC for the EMCI vs LMCI and LMCI vs AD classification. The below chance level accuracy and AUC in some cases is obtained due to the low sample size and high noise in the data, which is typical in neuroscience datasets (Jamalabadi et al., 2016). The performance is particularly poor for the NC vs EMCI classification as the changes captured by diffusion MRI based structural connectome networks are subtle (Figure 6, row 2). Incorporating feature engineering and/or feature selection may help further improve the classification performance at the cost of model interpretability.

Logistic regression allows easy interpretation of the coefficients in terms of odds ratios (Section 2.1.3). After training the models, top 10 coefficients with highest odds ratios were chosen to analyze the connectivity patterns that best explain disease transition. Figure 7 and Figure 8 show the brain connectivity patterns for EMCI vs LMCI and LMCI vs AD classifications, respectively. These connectivity maps were generated using BrainNet Viewer (Xia et al., 2013) (http://www.nitrc.org/projects/bnv/). Connectivity patterns were not analyzed for the NC vs EMCI case due to near chance level performance of the classifier. Abnormal inter-hemispherical connectivity patterns are evident in the EMCI vs LMCI classification. This could be due to atrophy in the corpus callosum associated with Mild Cognitive Impairment (Elahi et al., 2015). Abnormal connectivity patterns on the right side of the brain are involved in the later stages of disease progression between LMCI and AD which warrant further investigation to validate the biological

basis of the observed connectivity patterns. The details of top 10 edges with highest odds ratios associated with EMCI vs LMCI and LMCI vs AD are in Table 7 and Table 8, respectively.

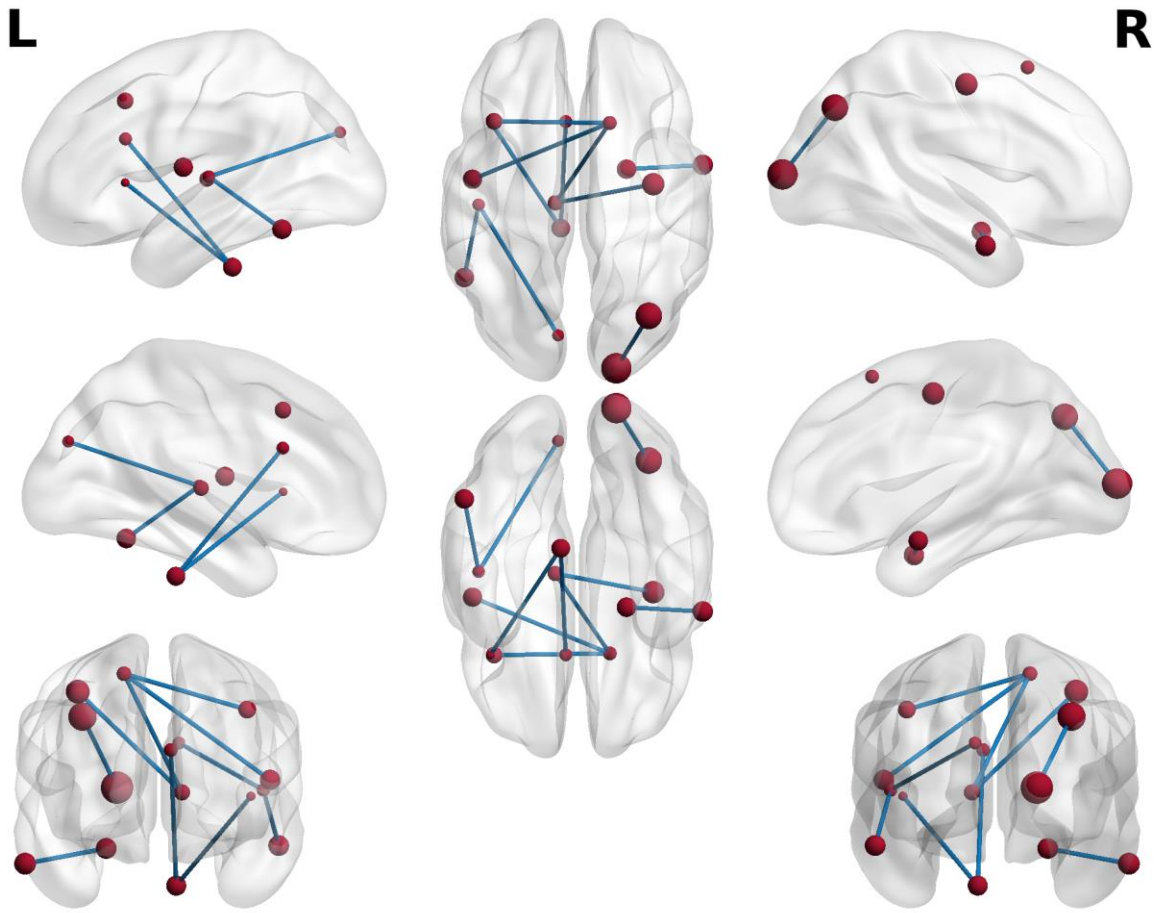

Figure 7: EMCI vs LMCI: Top 10 edges associated with highest odds ratio obtained from SSL

Table 7: Top 10 edges associated with highest odds ratio in EMCI vs LMCI classification

Sr. No.	Edge
1	"right lateral occipital cortex inferior division - right occipital pole"
2	"left thalamus - right superior frontal gyrus"
3	"left thalamus - right precentral gyrus"
4	"left frontal opercular cortex - left cerebellum"
5	"left cuneal cortex - left heschl"
6	"right superior frontal gyrus - left middle frontal gyrus"
7	"right amygdala - right middle temporal gyrus anterior division"
8	"left inferior temporal gyrus temporooccipital part - left heschl"
9	"left cingulate gyrus anterior division - left cerebellum"
10	"right superior frontal gyrus - left central opercular cortex"

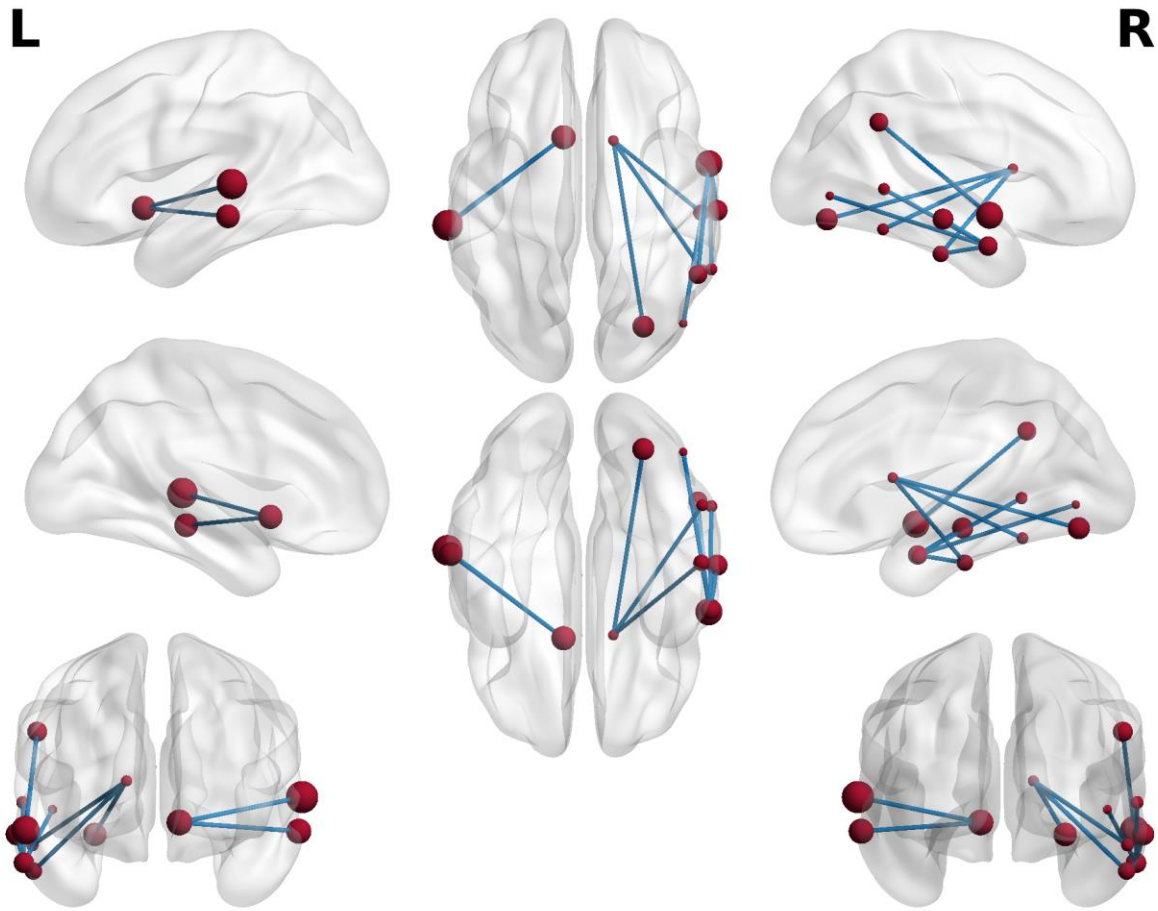

Figure 8 LMCI vs AD: Top 10 edges associated with highest odds ratio obtained from SSL

Table 8: Top 10 edges associated with highest odds ratio in LMCI vs AD classification

Sr. No.	Edge
1	"right superior temporal gyrus anterior division - right angular gyrus"
2	"right middle temporal gyrus anterior division - right middle temporal gyrus posterior division"
3	"left accumbens - left superior temporal gyrus posterior division"
4	"right middle temporal gyrus anterior division - right lateral occipital cortex inferior division"
5	"right middle temporal gyrus anterior division - right middle temporal gyrus temporooccipital part"
6	"right middle temporal gyrus anterior division - right inferior temporal gyrus posterior division"
7	"right caudate - right occipital fusiform cortex"
8	"right caudate - right inferior temporal gyrus posterior division"
9	"right caudate - right inferior temporal gyrus temporooccipital part"
10	"left accumbens - left middle temporal gyrus posterior division"

5.0 Conclusion and Future Scope

This thesis proposes a new framework for Shared Simultaneous Learning of logistic regression classifiers that allows information sharing between multiple binary classifiers during the training. A new term was added to the logistic log-likelihood function that constrains the K weight vectors associated with K binary classifications to be similar to each other by minimizing the squared Euclidean distance between them. Experimental results on diffusion MRI derived structural brain connectome from the ADNI2 dataset showed that SSL improved model performance as compared to the baseline independently trained multiple binary classifiers. SSL achieved an AUC of 0.53 ± 0.13 for NC vs EMCI, 0.68 ± 0.11 for EMCI vs LMCI and 0.73 ± 0.12 for LMCI vs AD classification. The marginal improvements in performance could be due to the small sample size and high dimensionality of the dataset under use. Nevertheless, this work highlights the advantage of information sharing across multiple classifiers during training. Further experiments using large datasets could be conducted to validate the proposed approach. Other differentiable distance or similarity metrics can also be used instead of the squared Euclidean norm for information sharing. One promising metric to test would be to maximize the cosine similarity between the weight vectors as it is more stable in higher dimensions. Also, the proposed approach may not be suitable when the number of classes is large as it may introduce a large bias.

After training the models, top 10 features with highest odds ratios were identified to analyze the connectivity patterns associated with Alzheimer's Disease progression. While the connectivity patterns for NC vs EMCI may not be informative due to chance level model performance, impaired inter-hemispherical connectivity patterns were identified in the EMCI vs LMCI classification and abnormal connectivity patterns were evident in the right hemisphere of

the brain for the LMCI vs AD classification. It must be noted that a different sample of the training data might result in a different set of connectivity patterns for the classifiers. Thus, a scoring approach such as that employed by (Q. Wang et al., 2018b) may further improve the stability of the results.

Bibliography

2021 Alzheimer's disease facts and figures. (2021). *Alzheimer's and Dementia*, *17*(3), 327–406. https://doi.org/10.1002/alz.12328

Andersson, J. L. R., & Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *NeuroImage*, *125*, 1063–1078. https://doi.org/10.1016/j.neuroimage.2015.10.019

Bassett, D. S., & Bullmore, E. T. (2009). Human brain networks in health and disease. In *Current Opinion in Neurology* (Vol. 22, Issue 4, pp. 340–347). NIH Public Access. https://doi.org/10.1097/WCO.0b013e32832d93dd

Billeci, L., Badolato, A., Bachi, L., & Tonacci, A. (2020). Machine learning for the classification of alzheimer's disease and its prodromal stage using brain diffusion tensor imaging data: A systematic review. In *Processes* (Vol. 8, Issue 9, p. 1071). MDPI AG. https://doi.org/10.3390/pr8091071

Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. In *Nature Reviews Neuroscience* (Vol. 10, Issue 3, pp. 186–198). Nature Publishing Group. https://doi.org/10.1038/nrn2575

Chen, D. H., Zhang, L., & Ma, C. (2020). A Multimodal Diagnosis Predictive Model of Alzheimer's Disease with Few-shot Learning. *Proceedings - 2020 International Conference on Public Health and Data Science, ICPHDS 2020*, 273–277. https://doi.org/10.1109/ICPHDS51617.2020.00060

Chételat, G. (2018). Multimodal Neuroimaging in Alzheimer's Disease: Early Diagnosis, Physiopathological Mechanisms, and Impact of Lifestyle. In *Journal of Alzheimer's Disease* (Vol. 64, Issue s1, pp. S199–S211). IOS Press. https://doi.org/10.3233/JAD-179920

Damoiseaux, J. S. (2012). Resting-state fMRI as a biomarker for Alzheimer's disease. In *Alzheimer's Research and Therapy* (Vol. 4, Issue 3, pp. 1–2). BioMed Central. https://doi.org/10.1186/alzrt106

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Dickerson, B. C., Stoub, T. R., Shah, R. C., Sperling, R. A., Killiany, R. J., Albert, M. S., Hyman, B. T., Blacker, D., & Detoledo-Morrell, L. (2011). Alzheimer-signature MRI biomarker

predicts AD dementia in cognitively normal adults. *Neurology*, *76*(16), 1395–1402. https://doi.org/10.1212/WNL.0b013e3182166e96

Ebrahimighahnavieh, M. A., Luo, S., & Chiong, R. (2020). Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. *Computer Methods and Programs in Biomedicine*, *187*, 105242. https://doi.org/10.1016/j.cmpb.2019.105242

Elahi, S., Bachman, A. H., Lee, S. H., Sidtis, J. J., & Ardekani, B. A. (2015). Corpus Callosum Atrophy Rate in Mild Cognitive Impairment and Prodromal Alzheimer's Disease. *Journal of Alzheimer's Disease*, *45*(3), 921–931. https://doi.org/10.3233/JAD-142631

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

FDA's Decision to Approve New Treatment for Alzheimer's Disease | FDA. (n.d.). Retrieved June 26, 2021, from https://www.fda.gov/drugs/news-events-human-drugs/fdas-decision-approve-new-treatment-alzheimers-disease

Feng, Q., & Ding, Z. (2020). MRI Radiomics Classification and Prediction in Alzheimer's Disease and Mild Cognitive Impairment: A Review. *Current Alzheimer Research*, *17*(3), 297–309. https://doi.org/10.2174/1567205017666200303105016

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, *54*(1), 313–327. https://doi.org/10.1016/j.neuroimage.2010.07.033

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102. https://doi.org/10.1016/s1053-8119(09)70884-5

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Hoerl, A. E., & Kennard, R. W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *42*(1), 80. https://doi.org/10.2307/1271436

Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W., & Evans, A. C. (1998). Enhancement of MR images using registration for signal averaging. *Journal of Computer Assisted Tomography*, *22*(2), 324–333. https://doi.org/10.1097/00004728-199803000-00032

Iglesias, J. E., Liu, C. Y., Thompson, P. M., & Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, *30*(9), 1617–1634. https://doi.org/10.1109/TMI.2011.2138152

Jamalabadi, H., Alizadeh, S., Schönauer, M., Leibold, C., & Gais, S. (2016). Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked

statistical properties of linear parametric classifiers. *Human Brain Mapping*, *37*(5), 1842–1855. https://doi.org/10.1002/hbm.23140

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156. https://doi.org/10.1016/S1361-8415(01)00036-6

Jo, T., Nho, K., & Saykin, A. J. (2019). Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Frontiers in Aging Neuroscience*, *11*, 220. https://doi.org/10.3389/fnagi.2019.00220

Kim, Y., Jiang, X., Giancardo, L., Pena, D., Bukhbinder, A. S., Amran, A. Y., & Schulz, P. E. (2020). Multimodal Phenotyping of Alzheimer's Disease with Longitudinal Magnetic Resonance Imaging and Cognitive Function Data. *Scientific Reports*, *10*(1), 1–10. https://doi.org/10.1038/s41598-020-62263-w

Kocahan, S., & Doğan, Z. (2017). Mechanisms of Alzheimer's disease pathogenesis and prevention: The brain, neural pathology, N-methyl-D-Aspartate receptors, tau protein and other risk factors. In *Clinical Psychopharmacology and Neuroscience* (Vol. 15, Issue 1, pp. 1–8). Korean College of Neuropsychopharmacology. https://doi.org/10.9758/cpn.2017.15.1.1

Lazar, M., Weinstein, D. M., Tsuruda, J. S., Hasan, K. M., Arfanakis, K., Meyerand, M. E., Badie, B., Rowley, H. A., Haughton, V., Field, A., & Alexander, A. L. (2003). White matter tractography using diffusion tensor deflection. *Human Brain Mapping*, *18*(4), 306–321. https://doi.org/10.1002/HBM.10102

Ledig, C., Schuh, A., Guerrero, R., Heckemann, R. A., & Rueckert, D. (2018). Structural brain imaging in Alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. *Scientific Reports*, *8*(1), 11258. https://doi.org/10.1038/s41598-018-29295-9

Lee, S., Lee, H., Abbeel, P., & Ng, A. Y. (2006). Efficient L1 Regularized Logistic Regression. *In: AAAI*.

Leow, A. D., Yanovsky, I., Chiang, M. C., Lee, A. D., Klunder, A. D., Lu, A., Becker, J. T., Davis, S. W., Toga, A. W., & Thompson, P. M. (2007). Statistical properties of Jacobian maps and the realization of unbiased large-deformation nonlinear image registration. *IEEE Transactions on Medical Imaging*, *26*(6), 822–832. https://doi.org/10.1109/TMI.2007.892646

Li, Y., Wang, S., Tian, Q., & Ding, X. (2014). Learning cascaded shared-boost classifiers for part-based object detection. *IEEE Transactions on Image Processing*, *23*(4), 1858–1871. https://doi.org/10.1109/TIP.2014.2307432

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., & Feng, D. (2014). Early diagnosis of Alzheimer's disease with deep learning. *2014 IEEE 11th International Symposium on Biomedical Imaging, ISBI 2014*, 1015–1018. https://doi.org/10.1109/isbi.2014.6868045

Márquez, F., & Yassa, M. A. (2019). Neuroimaging Biomarkers for Alzheimer's Disease. In *Molecular Neurodegeneration* (Vol. 14, Issue 1, pp. 1–14). BioMed Central Ltd. https://doi.org/10.1186/s13024-019-0325-5

McCullagh, P., & Nelder, J. A. (1983). *Generalized Linear Models* (2nd ed.). Routledge. https://doi.org/10.1201/9780203753736

Mitchell, A. J., & Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia - Meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica*, *119*(4), 252–265. https://doi.org/10.1111/j.1600-0447.2008.01326.x

Molnar, C. (2019). *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/

Petersen, R. C. (2011). Mild cognitive impairment. *The New England Journal of Medicine*, *364*(23), 2227–2234. https://doi.org/10.1056/NEJMcp0910237

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology*, *56*(3), 303–308. https://doi.org/10.1001/archneur.56.3.303

Rodríguez, G. (2007). *Lecture Notes on Generalized Linear Models*. https://data.princeton.edu/wws509/notes/

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, *52*(3), 1059–1069. https://doi.org/10.1016/j.neuroimage.2009.10.003

Sánchez-Maroño, N., Alonso-Betanzos, A., García-González, P., & Bolón-Canedo, V. (2010). Multiclass classifiers vs multiple binary classifiers using filters for feature selection. *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN.2010.5596567

Shalev-Shwartz, S., Wexler, Y., & Shashua, A. (2011). ShareBoost: Efficient Multiclass Learning with Feature Sharing. *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 1179–1187.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. https://doi.org/10.1002/hbm.10062

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., de Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., de Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in

functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*(SUPPL. 1), S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Sperling, R. (2011). The potential of functional MRI as a biomarker in early Alzheimer's disease. *Neurobiology of Aging*, *32*(SUPPL. 1), S37. https://doi.org/10.1016/j.neurobiolaging.2011.09.009

Teipel, S., Drzezga, A., Grothe, M. J., Barthel, H., Chételat, G., Schuff, N., Skudlarski, P., Cavedo, E., Frisoni, G. B., Hoffmann, W., Thyrian, J. R., Fox, C., Minoshima, S., Sabri, O., & Fellgiebel, A. (2015). Multimodal imaging in Alzheimer's disease: Validity and usefulness for early detection. In *The Lancet Neurology* (Vol. 14, Issue 10, pp. 1037–1053). Lancet Publishing Group. https://doi.org/10.1016/S1474-4422(15)00093-9

The Brain Imaging Software Toolbox. (n.d.). Retrieved June 29, 2021, from http://www.bic.mni.mcgill.ca/software/

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tournier, J. D. (2019). Diffusion MRI in the brain – Theory and concepts. In *Progress in Nuclear Magnetic Resonance Spectroscopy* (Vols. 112–113, pp. 1–16). Elsevier B.V. https://doi.org/10.1016/j.pnmrs.2019.03.001

Wang, K., Liang, M., Wang, L., Tian, L., Zhang, X., Li, K., & Jiang, T. (2007). Altered functional connectivity in early Alzheimer's disease: A resting-state fMRI study. *Human Brain Mapping*, *28*(10), 967–978. https://doi.org/10.1002/hbm.20324

Wang, Q., Guo, L., Thompson, P. M., Jack, C. R., Dodge, H., Zhan, L., & Zhou, J. (2018a). The Added Value of Diffusion-Weighted MRI-Derived Structural Connectome in Evaluating Mild Cognitive Impairment: A Multi-Cohort Validation. *Journal of Alzheimer's Disease*, *64*(1), 149–169. https://doi.org/10.3233/JAD-171048

Wang, Q., Guo, L., Thompson, P. M., Jack, C. R., Dodge, H., Zhan, L., & Zhou, J. (2018b). The Added Value of Diffusion-Weighted MRI-Derived Structural Connectome in Evaluating Mild Cognitive Impairment: A Multi-Cohort Validation. *Journal of Alzheimer's Disease*, *64*(1), 149–169. https://doi.org/10.3233/JAD-171048

Wang, R., Benner, T., Sorensen, A. G., & Wedeen, V. J. (2007). Diffusion Toolkit: A Software Package for Diffusion Imaging Data Processing and Tractography. *Proc Intl Soc Mag Reson Med*, *15*(3720).

Xia, M., Wang, J., & He, Y. (2013). BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics. *PLoS ONE*, *8*(7), 68910. https://doi.org/10.1371/journal.pone.0068910

Zhan, L., Mueller, B. A., Jahanshad, N., Jin, Y., Lenglet, C., Yacoub, E., Sapiro, G., Ugurbil, K., Harel, N., Toga, A. W., Lim, K. O., & Thompson, P. M. (2013). Magnetic resonance field

strength effects on diffusion measures and brain connectivity networks. *Brain Connectivity*, *3*(1), 72–86. https://doi.org/10.1089/brain.2012.0114

Zhan, L., Zhou, J., Wang, Y., Jin, Y., Jahanshad, N., Prasad, G., Nir, T. M., Leonardo, C. D., Ye, J., & Thompson, P. M. (2015). Comparison of nine tractography algorithms for detecting abnormal structural brain networks in Alzheimer's disease. *Frontiers in Aging Neuroscience*, *7*(APR), 48. https://doi.org/10.3389/fnagi.2015.00048

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zweig, A., & Weinshall, D. (2013). Hierarchical Regularization Cascade for Joint Learning. *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, III–37–III–45.