Machine Learning Methods for Biosignature Discovery

by

Rashmi Dubey

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2012 by the
Graduate Supervisory Committee:

Jieping Ye, Chair
Yalin Wang
Tong Wu

ARIZONA STATE UNIVERSITY

August 2012

ABSTRACT

Alzheimer′s Disease (AD) is the most common form of dementia observed in elderly patients and has significant social-economic impact. There are many initiatives which aim to capture leading causes of AD. Several genetic, imaging, and biochemical markers are being explored to monitor progression of AD and explore treatment and detection options. The primary focus of this thesis is to identify key biomarkers to understand the pathogenesis and prognosis of Alzheimer′s Disease.

Feature selection is the process of finding a subset of relevant features to develop efficient and robust learning models. It is an active research topic in diverse areas such as computer vision, bioinformatics, information retrieval, chemical informatics, and computational finance. In this work, state of the art feature selection algorithms, such as Student′s t-test, Relief-F, Information Gain, Gini Index, Chi-Square, Fisher Kernel Score, Kruskal-Wallis, Minimum Redundancy Maximum Relevance, and Sparse Logistic regression with Stability Selection have been extensively exploited to identify informative features for AD using data from Alzheimer′s Disease Neuroimaging Initiative (ADNI). An integrative approach which uses blood plasma protein, Magnetic Resonance Imaging, and psychometric assessment scores biomarkers has been explored. This work also analyzes the techniques to handle unbalanced data and evaluate the efficacy of sampling techniques. Performance of feature selection algorithm is evaluated using the relevance of derived features and the predictive power of the algorithm using Random Forest and Support Vector Machine classifiers. Performance metrics such as Accuracy, Sensitivity and Specificity, and area under the Receiver Operating Characteristic curve (AUC) have been used for evaluation. The feature selection algorithms best suited to analyze AD proteomics data have been proposed. The key biomarkers distinguishing healthy and AD patients, Mild Cognitive Impairment (MCI) converters and non-converters, and healthy and MCI patients have been identified.

To Dear Vivek

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to my adviser, Professor Jieping Ye, for his constant support, encouragement and valuable suggestions.

My sincere thanks go to the members of my dissertation committee, Professor Yalin Wang and Professor Teresa Wu for their thoughtful comments.

I am also thankful to my fellow members in the Center for Evolutionary Medicine & Informatics (CEMI) at Arizona State University: Arun Buduru, Rita Chattopadhyay, Jianhui Chen, Pinghua Gong, Shuiwang Ji, Bin Bin Lin, Yashu Liu, Zhi Nie, Cheng Pan, Zhisong Pang, Liang Sun, Qian Sun, Ramesh Thulasiram, Jie Wang, Zheng Wang, Carol Williams, Jason Wolf, Shuo Xiang, Lei Yuan, Sen Yang, Chao Zhang, Lei Zhang, and Jiayu Zhou. Thanks also are due to my friends for making my stay in Tempe such an enjoyable experience.

I am most indebted to my husband, my parents and parents-in-law, my brothers and sisters, and relatives for their constant support and encouragement throughout the course of my education.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Motivation

Alzheimer′s disease (AD) is the most frequent form of dementia in elderly patients. It destroys brain cells resulting in loss of memory, and cognitive and behavioral problems. With the advent of modern medicine, the average life expectancy has been on the rise. The percentage of population that is elderly is increasing. The effects of AD on the patient are debilitating; it also has profound adverse impacts on our social fabric in addition to the economic impact on families. There is currently no cure for AD: It is imperative that we understand the origins of this disease and strive towards mitigating the risk for future generations.

The root cause for AD is unknown; however the importance of genes in the onset and development of this disease is unquestioned. Early onset of AD in most cases is inherited and has been linked to any one of a number of different single-gene mutations on particular chromosomes. These mutations cause formation of abnormal proteins, for example, abnormal amyloid precursor protein (APP). Imaging techniques have been developed to study the accumulation of amyloid in the brain and its correlation to genetics. Late-onset AD is likely caused by a combination of genetic, lifestyle and environmental factors. A genetic risk factor has been identified which appears to increase the risk for developing this disease. Apolipoprotein E (APOE) gene comes in different forms or alleles. People with an APOE $\varepsilon 4$ allele are likely to develop Alzheimer′s Disease than those who do not have an APOE $\varepsilon 4$ [1].

Genome wide association studies have been able to identify a number of genes which may increase the likelihood of a person developing AD. The importance of continuing research in this area cannot be understated. As genetic research identifies more risk-factors, genetic testing will become critical in identifying populations at risk. As therapies to mitigate onset and progression of AD become available, it will be important to identify risk populations early before clinical symptoms appear. In addition, genetic

studies correlating environmental factors and expression of genes, also known as epigenetics, will be critical in minimizing the spread of this disease.

1.2 Problem Statement

There are many initiatives which aim to capture leading causes of AD. Several genetic, imaging and biochemical markers are being explored to monitor progression of AD and explore treatment and detection options. The complex nature of AD calls for a multipronged approach which looks at a multitude of factors as any single factor is unlikely to be an effective biomarker. Alzheimer′s Disease Neuroimaging Initiative (ADNI), a collaborative effort by multiple research groups was launched in 2004 to help identify the combination of biomarkers with the highest diagnostic and prognostic power. This initiative has helped develop optimized methods and uniform standards for acquiring biomarker data which includes Magnetic Resonance Imaging (MRI), positron emission tomography with Pittsburgh Compound B (PiB PET imaging), proteomics (blood protein) data, genetic variants amongst population such as Single Nucleotide Polymorphism (SNP), structural and metabolic imaging data on patients with AD, Mild Cognitive Impairment (MCI) and healthy controls, and creating an accessible data repository for the scientific community [36]. An illustrative example depicting the brain cell deterioration is shown in Figure 1.1. A key issue encountered during the search for the most potent biomarkers is measurement technique related variability and variability due to subject-related factors. Development of optimized and standardized techniques helps address the measurement-technique related variability. Subject-related factors can be multifactorial such as genetic predisposition, concurrent diseases and their treatment and effect of AD treatments. The overarching goal is to diagnose the disease in the earliest stage so that available therapies can be used to decelerate its progression.

The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) is a study to discover factors which determine subsequent development of symptomatic Alzheimer′s Disease [2]. It was launched in 2006 and over a span of 4.5 years sought to

study 1000+ participants which included AD and MCI patients and healthy volunteers. It used a multidisciplinary approach with 4 research streams to investigate cognitive, imaging, biomarkers and lifestyle factors. Volunteers underwent a screening interview, had comprehensive cognitive testing, gave 80 ml of blood, and completed health and lifestyle questionnaires. One quarter of the sample also underwent amyloid PET brain imaging with Pittsburgh compound B (PiB PET) and MRI brain imaging, and a subgroup of 10% had ActiGraph activity monitoring and body composition scanning [15]. The AIBL seeks to provide a unique data repository and sample population for study of AD which is continuously being re-assessed to determine the predictive efficacy of biomarkers, cognitive and lifestyle factors. The focus of this work is ADNI data.

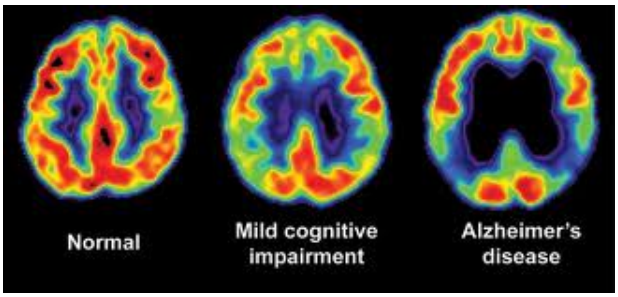


Figure 1.1: PET Scan of Normal, MCI, and AD patients [3]

The genome consists of the organism′s entire hereditary information including both the genes and the non-coding sequences of the DNA and RNA. Genomics is a discipline in genetics concerned with the study of the genomes of organisms. Proteins play a critical role in living organisms as the main components of the metabolic pathways of cells. Proteomics is the study of proteins, in particular their structures and functions. The proteome is the entire complement of proteins including modifications made to the protein by the organism or the system. The study of proteomes is more complex than genomics because unlike the organism′s genome which is constant, proteome differs from cell to cell and with time.

Considering the rapid growth in population of the elderly, it is imperative that a

multi-stage approach be used for screening and diagnosis. Blood-based biomarkers have the potential to offer significant advantages as a diagnostic tool as they are a cost-effective method of screening patients which can be followed up by more complex tests such as neuropsychological testing, neuroimaging, etc. This research work primarily focuses on identifying blood plasma proteins signatures. Additionally, informative genes and MRI features have also been identified which can help in AD prognosis.

Microarrays are tools for analyzing large number of gene or protein expressions on a small glass slide. This path-breaking technique helps in analyzing and comparing large samples quickly, for example, studying gene expression patterns between healthy and diseased tissues. Understanding the differences in gene and protein patterns between healthy and diseased samples is the critical first step in developing techniques to diagnose diseases, monitor their progression and identify treatment options. The data collected using microarray assays is analyzed using various statistical techniques such as clustering, classification and density estimation. A key challenge for these statistical techniques is the presence of small number of observations while the number of features present is enormous. This calls for use of advanced feature selection algorithms.

The famous German physicist, Albert Einstein, once said *Information is not knowledge*. This adage aptly fits the current scenario where acquiring and storing information is getting cheaper day by day, but it is becoming increasingly difficult to get meaningful knowledge from that information. If the ratio of features to size of training samples is very high, learning models tend to over-fit the training set and fail to generalize to unknown test data. Often the amount of training samples required to statistically analyze such multi-dimensional data grows exponentially with the dimensionality, a phenomenon commonly known as Curse of Dimensionality. Feature selection, also known as dimensionality reduction, is a technique to reduce the dimensionality of the data by retaining only relevant features for the statistical analysis. It is a data pre-processing step before building a learning model. The benefits of feature selection are many fold. Apart from alleviating the curse of dimensionality, it helps in reducing the noisy variables

which can otherwise misrepresent the data. Even in the absence of noisy variables, it is beneficial to remove redundant or irrelevant features for the sake of model interpretability and efficiency of the algorithm. Extracting informative features based on the data is a delicate task which must be done carefully to obtain robust results.

1.3 Background and Related Work

Several research studies have been focused on the use of feature selection algorithms to analyze AD data. Ray *et al.* [39] identified 18 signaling proteins in blood plasma which can classify Alzheimer′s Patients and Normal Control subjects with nearly 90% accuracy. They were also able to predict the mild cognitive impairment patients who may convert to AD in near future. Ravetti and Moscato [38] analyzed molecular data using a four step process which included abundance quantization, feature selection, literature analysis and selection of a classifier algorithm which is independent of the feature selection process. Their study resulted in the identification of a 5-protein biomarker molecular signature that achieves, on an average, a 96% total accuracy in predicting clinical AD. In a recent study on ADNI data by Daniel *et al.* [25], 11 protein signatures were identified which can predict pre-clinical AD achieving values of specificity and sensitivity between 65% and 86%. They used a multivariate approach based on combinatorial optimization ((α,β)-*k Feature Set Selection*), which retains information about individual participants and maintains the context of interrelationships between different features, to identify the optimal set of biomarkers.

Traditional gene selection methods do not account for the high degree of redundancy among the top ranked genes for a particular dataset. Presence of redundant genes can reduce the representativeness of a selected gene for the targeted classes and also increase the dimensionality of the selected genes adversely impacting the classification/prediction performance. The input data can have redundant information, in other words, correlated or dependent features. No additional information is gained when these redundant features are present together. Sometimes features which are

not of any use by themselves can prove to be very useful when combined with other features. Isabelle *at el.* have given a good discourse of relevant, redundant, correlated, and useless features [20]. Pair-wise Gene-to-gene correlation has been used by some researchers to remove redundant genes [14]. Some issues with this approach are the time complexity of $O(N^2)$ and the decision needed for the threshold for number of selected genes. Yu and Liu proposed an efficient way to tackle these issues by developing a novel method which does not require setting of a threshold in determining feature redundancy and combines sequential forward selection with elimination substantially reducing the number of feature pairs which need to be evaluated [49].

Tuv *et al.* [45] developed an algorithm using tree-based ensembles to generate a compact subset of non-redundant features. They utilized Random Forest algorithms which efficiently rank features for large data sets, and augmented the original data with artificial contrast variables which were constructed independent from the target and utilized their ranking for removing irrelevant variables from the original dataset. Serial ensembles were used to discover significant masking effects for redundancy elimination. Additionally, they employed an iterative strategy which allowed for weaker predictors to be identified after the stronger ones. Fung and Stoeckel [17] applied feature selection techniques for classification of SPECT images of Alzheimer′s disease. They used a linear programming formulation which incorporates proximity of features to generate a classifier which uses the most relevant areas for classification. Their study resulted in more robust classifiers with sensitivity considerably better than human experts. It is assumed that the real world data is generally Sparse, i.e. the set of most informative genes is very small. To take advantage of sparsity of the data, special machine learning techniques are required. Sparse Learning with Efficient Projection(SLEP) is a complete package of machine learning algorithms for finding sparse representations of data [21][32].

1.4 Thesis Organization

Chapter 2 gives the details of the feature selection algorithms used in generating the informative feature subset followed by model selection algorithms to measure the goodness of the former algorithm. Stability Selection approach, methods for handling unbalanced data, and performance metrics are also discussed. Chapter 3 gives the details of the datasets, experiments, and results obtained. Chapter 4 concludes the thesis with observations from current work and pointers to extend the work.

CHAPTER 2

SYSTEMS AND METHODOLOGY

This chapter lists the state-of-the-art techniques assayed in identifying potential biomarkers for Alzheimer′s Disease. The chapter starts with a brief discussion of various techniques commonly employed to prepare the data for analysis. A detailed description of various feature selection algorithms studied is summarized in section 2.3. Stability Selection, a procedure to ensure that the sub-feature space is stable, is described in section 2.4. A brief summary of classification techniques is listed in section 2.5.

2.1 Data Pre-Processing

Real world data is noisy and erroneous. Tools employed in data collection can have calibration issues which might result in missing or out of range values. Various steps are involved in preparing the raw data for the statistical analysis [29]. This section discusses few of the common measures taken to process the data.

2.1.1 Missing Values

Missing value is defined as the absence of the information about the variable in the given observation. Missing data can adversely impact the data analysis and result in erroneous conclusions. Imputation methods can be used to estimate missing values. Another approach is to discard the samples with missing values. The latter approach has been employed in this thesis.

2.1.2 Handling Unbalanced Data

This is one of the main issues in data classification. Unbalanced data refers to unequal number of observations in different classes, in other words, the observed data is biased towards the majority class. A number of approaches have been used to evaluate imbalanced data such as random over-sampling with replacement, random under-sampling,

and directed over-sampling. Sampling methods are often applied to the data used to build the learning model. The two sampling methods used in this work are discussed here.

2.1.2.1 Over Sampling

Over-sampling relies on random replication of minority samples to achieve balance in the class/category distribution. Over-sampling increases computational burden for large datasets.

2.1.2.2 Under Sampling

Random under-sampling aims to achieve balance by random elimination of some majority class samples. The downsides of this approach are the risk of discarding potentially important data and the loss of randomness of the sample distribution [28].

2.1.3 Feature Selection

There are many advantages of feature selection. It is one of the most fundamental data pre-processing steps. Section 2.3 discusses the commonly used feature selection methods.

2.1.4 Data Normalization

It is important to distinguish between natural variations in gene expression levels and variations induced by measurement noise. This technique is increasingly used to analyze microarray data to remove systematic effects and improve signal to noise ratio. Normalization helps pre-process the data to better suit the requirements of the algorithm. There are several techniques which can be employed to normalize the data. In mean normalization, the data is normalized so that the average of the distribution is zero. In unit normalization, the data is normalized such that the standard deviation is one. Generally, normalization techniques are used to make the values, which are measured using different scales, comparable. It is also known as *Feature Scaling*. Classification

algorithms like SVM using Gaussian Kernel require all features to be on same scale. For a given data distribution, $\boldsymbol{X} \in \mathbb{R}^{M \mathbf{X} N}$, feature scaling for i^{th} data point and j^{th} feature is computed using the formulae:

- Min-Max Normalization,

$$z_i^{(j)} = \frac{x_i^{(j)} - x_{min}^{(j)}}{x_{max}^{(j)} - x_{min}^{(j)}} \tag{2.1}$$

- Z-score Normalization,

$$z_i^{(j)} = \frac{x_i^{(j)} - \mu^{(j)}}{\sigma^{(j)}} \tag{2.2}$$

where $x_{min}^{(j)}$ is the minimum, $x_{max}^{(j)}$ is the maximum, $\mu^{(j)}$ is the mean and $\sigma^{(j)}$ is the standard deviation of the j^{th} feature vector.

2.2 Dataset Generation

The given data is partitioned into disjoint training and testing subsets to determine the efficacy and the generalization capability of the model learned. The learning model is built on training set and its power is determined on test or validation set. The train-test split ratio used is usually 7:3 or 9:1. There are various methods to generate train-test set. This section discusses the commonly used techniques.

2.2.1 Cross Validation

This is a common approach employed in classification problems. In this strategy, the data is repeatedly divided into training and test sets. Neither the training nor the testing sets should overlap in any of the iterations so that each data point is in the test set atleast once. The performance of the classifier is determined using various averaging techniques. In this work, simple averaging method is used. The following subsections give a brief account of the different cross validation approaches.

2.2.1.1 *k*-fold Cross Validation

The data is evenly partitioned into *k*-folds. A test set is generated using one of the folds and remaining *k*-1 folds are used as training data. This process is repeated *k* times, such that each of the *k* folds is used exactly once as the test set. This approach has been used in this thesis.

2.2.1.2 Leave-One-Out Cross-validation

This is a special case of *k*-fold cross validation, where *k*=N (N is the number of data points). Subsequently, the model is trained on $(N-1)$ data points and there are N rounds of training the model. This is indeed a computationally expensive procedure and is better suited for smaller datasets.

2.2.2 Random Subsampling

Random Subsampling is the most common approach to generate train-test datasets. Training data points are randomly selected and the remaining data points are used for test set. It is common to generate atleast 10 such datasets; these datasets can overlap and some data points will never be selected in either training or testing or both. The trick is to generate enough datasets, such that each data point is used atleast once. This approach is used in the thesis with 10 different train-test datasets using 9:1 train-test split ratio.

2.3 Feature Selection Algorithms

Feature selection (FS) is an important tool in Machine Learning. FS algorithms can be categorized into filter, wrapper, and embedded methods based on the criteria used for selecting the prominent features. Filter methods are based on ranking the features, usually analyzing each feature individually, using test statistics or by assigning them different weights depending upon their usefulness in separating the two classes. Due to the simplicity and deployment efficiency, filter methods are most commonly employed.

Wrapper methods are so called as they wrap any classification technique to select a sub-set of features by exhaustively searching the feature space. This technique can be computationally expensive with risk of overfitting the data. Embedded techniques are similar to wrapper methods, but unlike wrapper approach they use specific classification models. While most FS algorithms are designed to work on classification problems, some can be extended to perform regression analysis as well. In classification problems, the FS algorithms can further be classified into three categories: supervised, unsupervised, and semi-supervised, depending upon the input data, if the data is labeled, unlabeled or partially labeled respectively.

In this work, supervised classification based feature selection algorithms which utilize filter methods are examined. Features are assumed to be independent and identically distributed *(i.i.d)*, i.e. no redundant features. The matlab code for the FS algorithms is taken from ASU Feature Selection Repository [50] unless otherwise noted. This section gives a brief overview of various feature selection algorithms studied in this work.

2.3.1 Student's t*-test*

It is a statistical hypothesis test in which the test statistic follows a Student′s *t*-distribution if the null hypothesis, denoted by H_0, is supported. The alternative hypothesis, denoted by H_1, checks for the condition that H_0 does not hold. This test is suited for distributions which are smaller in size, are symmetric to normal distribution, and their variance is unknown. There are three types of *t*-test:

- One Sample *t*-test compares the sample mean with the known population mean or any fixed value. The null hypothesis states that the sample mean and the population mean or the given fixed value are equal.
- Unpaired *t*-test compares two samples which are independent and identically distributed. For example, one sample is drawn from the population of control subjects

and another sample is drawn from the population of subjects with illness. The null hypothesis states that the two samples have equal means and equal variances.

- Paired *t*-test compares two samples which are dependent. For example, the samples comprising same subjects before and after treatment. The null hypothesis states that the two samples have equal means.

Let X and Y be the samples drawn from a population, then the test statistic (t-score) is defined by the equation:

$$t = \frac{[\overline{x} - \overline{y}]}{SE} \tag{2.3}$$

where *SE*, the standard error of the sampling distribution is computed as:

$$SE = \sqrt{[(\frac{s_x^2}{n}) + (\frac{s_y^2}{m})]}$$

$\overline{x}$ and $\overline{y}$ are the sample means,n and m are the sample sizes, and s_x and s_y are the sample standard deviations of X and Y respectively. DF, the Degrees of freedom is given by the equation:

$$\text{DF} = \frac{[(\frac{s_x^2}{n}) + (\frac{s_y^2}{m})]^2}{[\frac{(\frac{s_x^2}{n})^2}{n-1}] + [\frac{(\frac{s_y^2}{m})^2}{m-1}]}$$

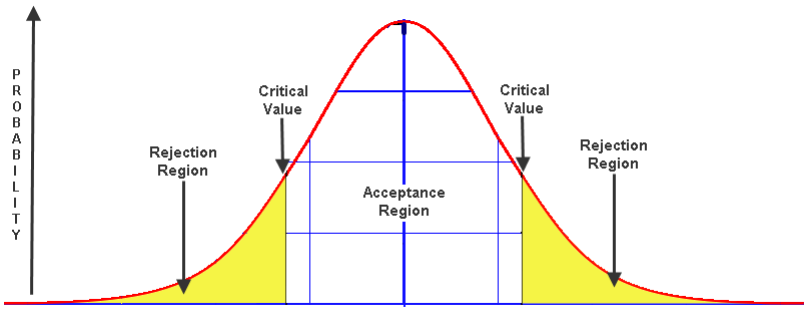


Figure 2.1: Acceptance and Rejection regions of a two-tailed Hypothesis Test. P-value lies in the Rejection region.

P-value is defined as the probability of observing a sample statistic as extreme or more extreme as test statistic under the null hypothesis. It is computed using a table of values from Student's *t*-distribution given a t-score and DF. The null hypothesis is

rejected if P-value is less than or equal to the significance level, denoted by $\alpha \leq$0.05. The features are ranked in ascending order of P-values. A one-tailed *t*-test tests the statistical significance in the one direction of interest according to the alternative hypothesis. This contrasts with a two-tailed *t*-test, in which the null hypothesis will be when the value of the test statistic is either sufficiently small or sufficiently large. In this work, unpaired, two-tailed *t*-test is used. The MathWork′s matlab *T-Test* function is used for this algorithm [19].

2.3.2 Relief-F

Relief-F is an extension of one of the most successful feature subset selection algorithms, Relief [26] based on relevance of features. The majority of feature selection algorithms estimate the quality of a feature based on its conditional independence upon the target class. Relief algorithm assesses the significance of a feature based on its ability to distinguish the neighboring instances. Let $\mathcal{X}$ be the sample distribution with M training data points in N dimensional feature space. To find the relevant feature subset of size $n \ll N$, Relief algorithm proceeds as follows:

(a) Initialize the weight vector $\mathcal{W}_{N\text{X}1}$ to zero.

(b) Select a triplet of data points: x_i, x_{same}, and x_{other}

where x_i is a randomly selected instance, x_{same} is a near-hit data point, a nearest neighbor which belongs to the same class as x_i, and x_{other} is a near-miss data point, a nearest neighbor which belongs to different class than x_i. N-dimensional euclidean distance is used to find x_{same} and x_{other}.

(c) The difference between a feature value from the two instances is computed as:

If the feature has continuous values:

$$\text{diff}(x_1(f), x_2(f)) = \frac{(x_1(f) - x_2(f))}{n_c}$$

where n_c is normalization unit to get diff value in [0,1] interval.

If the feature has categorical values:

$$\text{diff}(x_1(f), x_2(f)) = \begin{cases} 0, \text{ if } x_1(f) = x_2(f) \\ 1, \text{ otherwise} \end{cases}$$

(d) For every f^{th} feature:

$$\mathcal{W}(f) = \mathcal{W}(f) - \text{diff}(x_i(f), x_{same}(f))^2 + \text{diff}(x_i(f), x_{other}(f))^2 \qquad (2.4)$$

The underlying principle behind this equation is that if diff$(x_i(f), x_{same}(f))$ is large, then this feature f distinguishes data points within the same class. Such a feature is of no use and hence its weight should be reduced.

Whereas if diff$(x_i(f), x_{other}(f))$ is large, then this feature f distinguishes the data points from two different classes which serves the feature selection problem formulation well. Thus, the difference is added to increase the weight of such features.

(e) Repeat steps (b) and (d) for all instances of $\mathcal{X}$.

(f) Compute the average weight of each feature:

$$\mathcal{W}_{avg} = \frac{\mathcal{W}}{M}$$

(g) $\mathcal{W}_{avg}$ is sorted in descending order and top n features are returned as the relevant feature subset. Another way to extract the informative features is to provide a threshold τ such that any $\mathcal{W}_{avg}(f) < \tau$ will be discarded from the feature subset.

Relief-F algorithm improves Relief algorithm by introducing *k*-nearest neighbors from each class [40]. With this new approach, step(b) will give 2*k*+1 data points and in step(d), diff$(.,.)$ value is averaged over all *k* diff$(.,.)$ values. Choosing an optimal value of *k* is tricky; many values should be tried and the value which best generalizes the validation set should be used. A general rule of thumb is to start with *k*=10. This algorithm has been extended to handle incomplete data and multi-class problems [27].

2.3.3 Fisher Kernel Score

This is a comparatively new feature selection approach using Fisher Kernel [23] as the similarity measure between two data points. The key idea is to incorporate the goodness of generative and discriminative models. For example, a generative model, such as Naive Bayes or Hidden Markov Model, is trained on the data to generate a Fisher Kernel which is used in a kernel based discriminative model such as Support Vector Machine. Fisher Kernel is defined as:

$$K(x_i, x_j) = U_{x_i}^T I^{-1} U_{x_j} \tag{2.5}$$

where x_i and x_j are the data points, I is the Fisher Information Matrix, and U_x is the Fisher Score given by the equation:

$$U_x = \nabla_\theta log P(x|\theta)$$

where x is the data point, θ is the vector of parameters, $logP(x|\theta)$ is the log-likelihood of x with respect to the probability model given θ. ∇_θ is the gradient operator.

Jaakkola *et al.* [22] applied the method to detect remote protein homologies and showed the method works well in classifying protein domains by SCOP (Structural Classification of Proteins) superfamily. Fisher Kernel is more popular in web mining and speech recognition domains than in biomedical data analysis.

2.3.4 Gini Index

Gini Index, also known as Gini Coefficient or Gini Ratio, measures the inequality in the frequency distribution values. Its value lies between $[0, 1]$, where a coefficient of 0 indicates that all values of the distribution are the same and a coefficient of 1 indicates maximal inequality. This statistical measure of dispersion is commonly used to measure wealth or income inequality. It can be applied to various other fields as well. Mathematically it is defined as the ratio of the areas within the *Lorenz curve*. Gini Index is given

by the equation:

$$G = \frac{A}{(A+B)} \tag{2.6}$$

where A is the area above *Lorenz curve* and B is the area below it as shown in Figure. It can also be computed using the formula given by *Angus Deaton*:

$$G = \frac{N+1}{N-1} - \frac{2}{N(N-1)\mu}(\sum_{i=1}^{n} P_i X_i) \tag{2.7}$$

where N is the size of the population, μ is the mean, and P_i is the rank P of person i with income X.

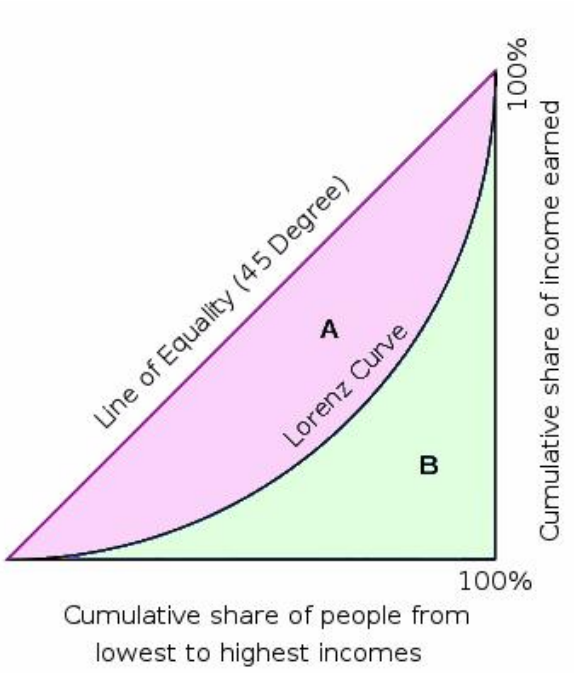


Figure 2.2: Graphical Representation of Gini Index [3]

2.3.5 Information Gain

Information Gain (IG) is also known as information divergence, Kullback-Leibler divergence, or relative entropy. Information gain is commonly used as a surrogate for approximating a conditional distribution in classification setting [13]. Let $\boldsymbol{X} \in \mathbb{R}^{M \mathbf{X} N}$ be the training set with M data points and N dimensions, and $\boldsymbol{Y} \in \mathbb{R}^{M \mathbf{X} 1}$ be the class label with c discrete values such that $\boldsymbol{Y} = \{y_1, y_2, ..., y_c\}$ and $y_i \neq y_j \forall i, j \in [1, c]$. Let $\boldsymbol{A} \in \mathbb{R}^{N \mathbf{X} 1}$ be the set of attributes of $\boldsymbol{X}$. An attribute $x_a \in \boldsymbol{A}$ can take k possible values

such that $x_a = \{v_1, v_2, ..., v_k\}$. Both $\boldsymbol{Y}$ and x_a are assumed to be discrete. IG measures the reduction in entropy in moving from a prior distribution $P(\boldsymbol{Y})$ to a posterior distribution $P(\boldsymbol{Y}|x_a)$. In simpler words, it represents the reduction in uncertainty of predicting $\boldsymbol{Y}$ given x_a. Information Gain is given by:

$$IG(\boldsymbol{Y}|x_a) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|x_a) \tag{2.8}$$

where $H(\boldsymbol{Y})$ and $H(\boldsymbol{Y}|x_a)$ are entropy functions defined as:

$$H(\boldsymbol{Y}) = -\sum_{i=1}^{c} P(\boldsymbol{Y} = y_i) log P(\boldsymbol{Y} = y_i)$$

$$H(\boldsymbol{Y}|x_a) = -\sum_{i=1}^{k} P(x_a = v_i) P(\boldsymbol{Y}|x_a = v_i)$$

An attribute with higher value of IG is considered more relevant and is assigned a higher weight. This is an asymmetric method, i.e. $IG(\boldsymbol{Y}|x_a) \neq IG(x_a|\boldsymbol{Y})$, and is not suitable for attributes which can take large number of discrete values (for example, when k is a large number) as it might cause overfitting problems.

2.3.6 Kruskal Wallis

Kruskal-Wallis is a non-parametric, rank based test to determine if the samples belong to the same distribution. It is based on one-way analysis of variance and does not assume a normal distribution. Its null hypothesis states that the populations from which the samples are drawn have same median. It can compare more than 2 groups, and is an extension of *Mann-Whitney U* test which analyzes a pair of sampling distributions. The most relevant feature is given a rank of 1, the second most relevant feature is ranked 2; similarly all N features are assigned ranks from 1 to N. Some information can get lost in the process of ranking the features [34]. The test statistic is given by:

$$K = (N-1)\frac{\sum_{i=1}^{g} n_i(\overline{r_i} - \overline{r})^2}{\sum_{i=1}^{g} n_i \sum_{j=1}^{n_i}(\overline{r_{ij}} - \overline{r})^2} \tag{2.9}$$

where n_i is the number of observations in the group i,

r_{ij} is the rank (among all observations) of observation j from group i,

N is the total number of observations across all groups,
$\overline{r_i} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$, and
$\overline{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.

Simplifying the denominator of Equation 2.9 we get:

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i(\overline{r_i})^2 - 3(N+1) \tag{2.10}$$

If the test statistic results are significant, then at least one sample is different from the rest samples; otherwise no difference exists between the samples. Repeated tests can be performed to exactly determine which of the two sample pairs are different. Due to high computational power required for the test, exact probabilities upto 105 samples are available [42][34].

2.3.7 Chi-Square Test

The Chi-Square (χ^2) test is a statistical test performed on samples that follow χ^2 distribution, a special case of *gamma distribution*. This distribution has the following property:

- It is continuous, asymmetrical, skewed to right distribution, and has k degrees of freedom such that $k = N - 1$, where N is the sample size.
- The mean of the distribution is equal to the degrees of freedom (k).
- The variance is twice the degrees of freedom ($2k$).
- It approaches normal distribution as value of k increases.
- The χ^2 distribution is widely used in χ^2 test to compute goodness of fit, independence of criteria, and estimating confidence interval and standard deviation.

In feature selection, χ^2 test for independence is employed to determine whether the outcome is dependent on a feature. The null hypothesis states that the occurrences

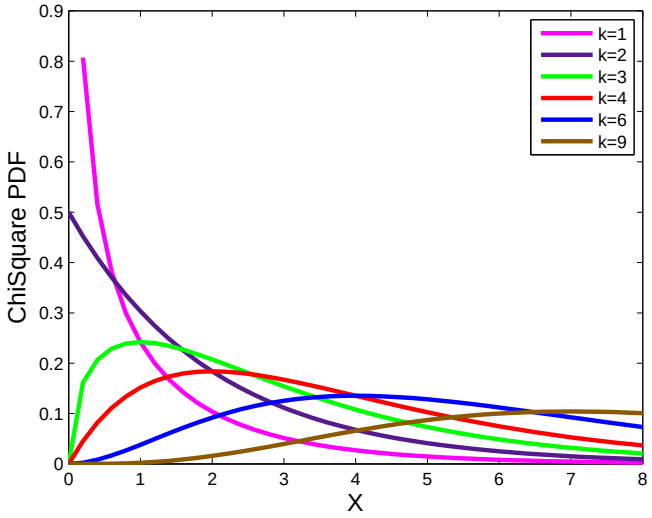


Figure 2.3: Probability Density Function for a family of χ^2 Distributions

of the outcomes of an observation are statistically independent. Given a feature with r distinct values and c possible outcomes, a contingency table is formed with r rows and c columns. Each cell (i, j) of the table, where $1 \leq i \leq r$ and $1 \leq j \leq c$, is filled with the observed frequency, O_{ij}, number of samples with i^{th} feature value and j^{th} outcome. The test statistic is defined as:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.11}$$

E_{ij} is the expected frequency given by:

$$E_{ij} = \frac{(\sum_{n_r=1}^{r} O_{n_r j})(\sum_{n_c=1}^{c} O_{i n_c})}{N}$$

where N is the sample size and the degrees of freedom $k = (r-1)(c-1)$. There must be atleast 5 observations per cell.

P-value is the probability of obtaining a test statistic as extreme as the observed value under null hypothesis. Given χ^2 test statistic and k, P-Value is computed using Chi-Square Distribution table. The null hypothesis is rejected if P-value is less than the specified significance level α, which is often ≤ 0.05.

Rejecting the hypothesis makes the result statistically significant and confirms the dependence of the outcome on the feature value.

2.3.8 Minimum-Redundancy-Maximum-Relevance (mRMR)

Minimum-Redundancy-Maximum-Relevance selection (mRMR) selects feature subset according to the maximum statistical dependency criterion based on mutual information as proposed by Peng *et al.* [37]. It selects the features which are highly correlated to the class variable (Maximum-Relevance), but are mutually far away from each other (Minimum-Redundancy). Maximum relevance of a feature subset S for a class c is given by

$$max\ D(S,c), D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c)$$

and, minimum redundancy condition of all the features in S is given by

$$min\ R(S), R = \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i;x_j)$$

The mRMR combines the two measures and optimizes them simultaneously:

$$max\ \Phi(D,R), \Phi = D - R$$

$$\Phi = \max_{s} \left[\ \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c) - \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i;x_j) \ \right] \tag{2.12}$$

where $I(x_i;c)$ or $I(x_i;x_j)$ represents the mutual information. For any two univariate random variable x and y, mutual information is defined in terms of their probabilistic density function as:

$$I(x;y) = \int\int p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy$$

Similarly, for multivariate random variables S_m and target class c:

$$I(S_m;c) = \int\int p(S_m,c) log \frac{p(S_m,c)}{p(S_m)p(c)} dS_m dc$$

Correlation score or distance/similarity scores can be used instead of mutual information. mRMR is a very robust approach but, in certain pathological scenarios, it fails

to evaluate the worth of the feature which is useless by itself but is very useful when combined with other features. This is due to inability to measure interaction between the features [9].

2.3.9 Sparse Logistic Regression

This FS algorithm is an example of filter method with regularization term. The guiding principle is to use regularization in Logistic loss function such that the informative features distinguishing the class are given more weight and other irrelevant features are given close to zero weight [31]. Let $f(\theta, b)$ be the logistic loss function, θ is the set of parameters or weight vector and b is the bias term. The ℓ_1-norm regularized logistic loss is defined as:

$$min \;\; f(\theta, b) + \frac{\lambda}{2}||\theta||_1 \tag{2.13}$$

where λ is the regularization parameter. ℓ_1-norm is known to induce sparsity [16]. The matlab code is taken from SLEP package [32]. Logistic Regression is discussed in detail in a subsequent section.

2.4 Stability Selection

It is a method based on a combination of subsampling and feature selection techniques to improve feature extraction and structure estimation [35]. A bootstrap sample, $\boldsymbol{X}_B \in \mathbb{R}^{K \mathrm{X} N}$, is obtained from the original data matrix $\boldsymbol{X} \in \mathbb{R}^{M \mathrm{X} N}$ using sampling with replacement technique. The approach is known as Bootstrap sampling. When $K = M$, the expected number of unique samples is given by:

$$(1 - \frac{1}{e}) \approx 63.2\%$$

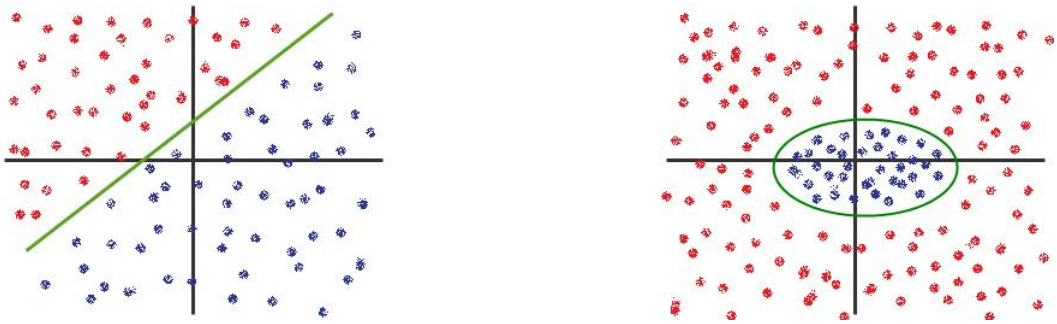
In practice, 500 to 1000 bootstrap samples are generated. Sparsity inducing feature selection method is applied to each bootstrap sample with a set of regularization parameters. Frequency of a feature being selected (i.e. its weight is not zero) in each bootstrap run is maintained for all regularization parameters. The maximum selection

frequency of the feature, also known as the probability of the feature over its stability path, is computed. Features are ranked in order of decreasing probability.

In this work, Sparse Logistic Regression and 1000 bootstrap samples (with $K = M$) are used with 10-12 regularization parameters such that the smallest regularization parameter selects 60% to 80% features and the largest regularization parameter selects 5% to 10% features.

2.5 Classification

Classification is a supervised learning method where a classifier is learned from the labeled training data and is used to identify the correct category of the new (unlabeled) instances. In binary classification data points belong to either of the two classes. In multiclass classification problems, the instance can belong to one of the several classes. This thesis focuses on binary classification.



(a) Linearly Separable (b) NonLinearly Separable

Figure 2.4: Different types of association between binomial data in 2-D plane

For a given training matrix $\boldsymbol{X} \in \mathbb{R}^{M \mathrm{X} N}$ and corresponding label vector $\boldsymbol{Y} \in \mathbb{R}^{M \mathrm{X} 1}$, the classification function is defined as:

$$f(\overrightarrow{\theta}, \overrightarrow{x})$$

where $\overrightarrow{\theta}$ is the set of parameters or weight vector, learned from labeled training data; $\overrightarrow{x}$ is a feature vector for the i^{th} sample; $f(\overrightarrow{\theta}, \overrightarrow{x})$ can be as simple a dot product or can be as complex as a kernel function.

Real world data can be linear or non-linear; Figure 2.4 illustrates these two cases in a 2-D plane. Classifiers can be divided into two categories: linear classifiers, which use linear combination of parameters and non-linear classifiers, which use non-linear combination of the parameters. Regularization term is often added to the objective function to prevent overfitting. This section discusses the commonly used classifiers.

2.5.1 Logistic Regression

Logistic Regression is a classification technique using linear discriminative model to maximize the quality of output on training data. For a two class (binomial) classification problem, it assigns a probability to class labels using the following logistic function:

$$h_\theta(x) = P(y = 1|x; \theta) = \frac{1}{1 + \exp^{-\theta^T x}} \tag{2.14}$$

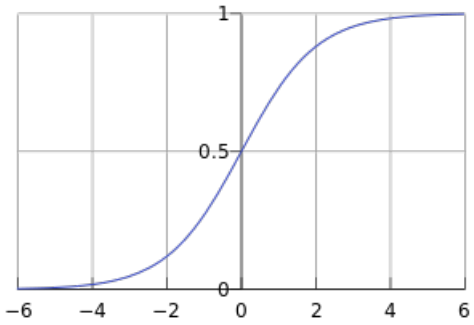


Figure 2.5: Standard Logistic (Sigmoid) Function Curve [3]

$h_\theta(x) \in [0, 1]$, is the estimated probability that the class label is positive, given x, and parameterized by θ. If $h_\theta(x) \geq 0.5$, the class label is positive otherwise it is negative, as illustrated in Figure 2.5. The cost (also known as logistic loss) function for any i^{th} sample is defined as:

$$\text{Cost}(h_\theta(x^{(i)}), y) = \begin{cases} -log(h_\theta(x^{(i)})) & if \;\; y = 1 \\ -log(1 - h_\theta(x^{(i)})) & if \;\; y = 0 \end{cases}$$

The objective function for Logistic Regression is defined as:

$$\min_{\theta} \ J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[\ y^{(i)}log(h_\theta(x^{(i)})) + (1-y^{(i)})log(1-h_\theta(x^{(i)})) \ \right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

where λ is the regularization parameter and $\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$ is the regularization term. In general, regularization term can be written as $\frac{\lambda}{2m}||\theta||_p$ where p represents the vector norm. ℓ_1-regularized norm is known to induce sparsity [16] and is preferred for sparse solutions. This approach can be extended to multiclass classification problems. In this work, ℓ_1-regularized norm and binomial Logistic Regression is used.

2.5.2 Support Vector Machine (SVM)

This is a supervised classification technique using discriminative linear models. Each data point is assigned to one and only one class. SVM constructs a hyperplane with the largest margin, hence it is also known as Large Margin or Maximum Margin Classifier. The equation of hyperplane dividing the two classes is given by:

$$\theta^T x - b = 0 \tag{2.15}$$

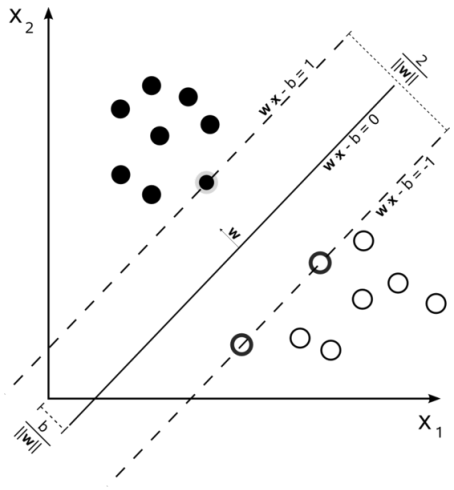
where $y_i \in \{-1, 1\}$, $\frac{b}{||\theta||}$ is the offset of hyperplane from the origin, $\overrightarrow{\theta}$ is perpendicular to the hyperplane, and θx is the dot product between θ and x. The equations of hyperplanes bounding the dividing hyperplane on either side are given by:

$$\theta^T x - b = 1 \text{ and } \theta^T x - b = -1 \tag{2.16}$$

The hyperplanes represented by equations 2.15 and 2.16 are illustrated in Figure 2.6. The distance between the hyperplanes defined in Equation 2.16 is $\frac{2}{||\theta||}$. In order to maximize this distance, $||\theta||$ must be minimized. This can be converted into a quadratic optimization problem and objective function of support vector machine is given by:

$$min\frac{||\theta||^2}{2}$$

subject to the constraint: $\quad y_i(\theta x_i - b) \geq 1 \quad$ for any $i, 1 \leq i \leq m$



(a) SVM Classifier

Figure 2.6: SVM Classifier in a 2-D plane; θ is represented by w here [3]

Large margin lowers the generalization error of the classifier [6]. Only a subset of training vectors, which lie on the hyperplanes given by Equation 2.16, are used to build the model. SVM can be employed for multiclass classification.

When the data is non-linearly separable in the given feature space as illustrated in Figure 2.4(b), kernel trick can be used to transform the data into a higher dimensional space so that new the data points are linearly separable in the new feature space. Non-linear kernel function is substituted for the dot product between θ and the data point x; it measures the similarity of each data point (x) in the original space to all landmark points (ℓ) in the new feature space and assigns the instance to the closest landmark. There are many types of kernel functions such as Polynomial, Gaussian, and Hyperbolic Tangent. In this work, Gaussian Radial Basis kernel function is used which is defined as:

$$k(x_i, \ell_j) = exp(-\frac{||x_i - \ell_j||^2}{2\sigma_2}) \tag{2.17}$$

Libsvm package [11] is used for the matlab implementation of SVM.

2.5.3 Random Forest (RF)

Random Forest (RF) is an ensemble classifier combining Bootstrap aggregation (bagging), random feature selection, and decision trees. RF grows many classification trees

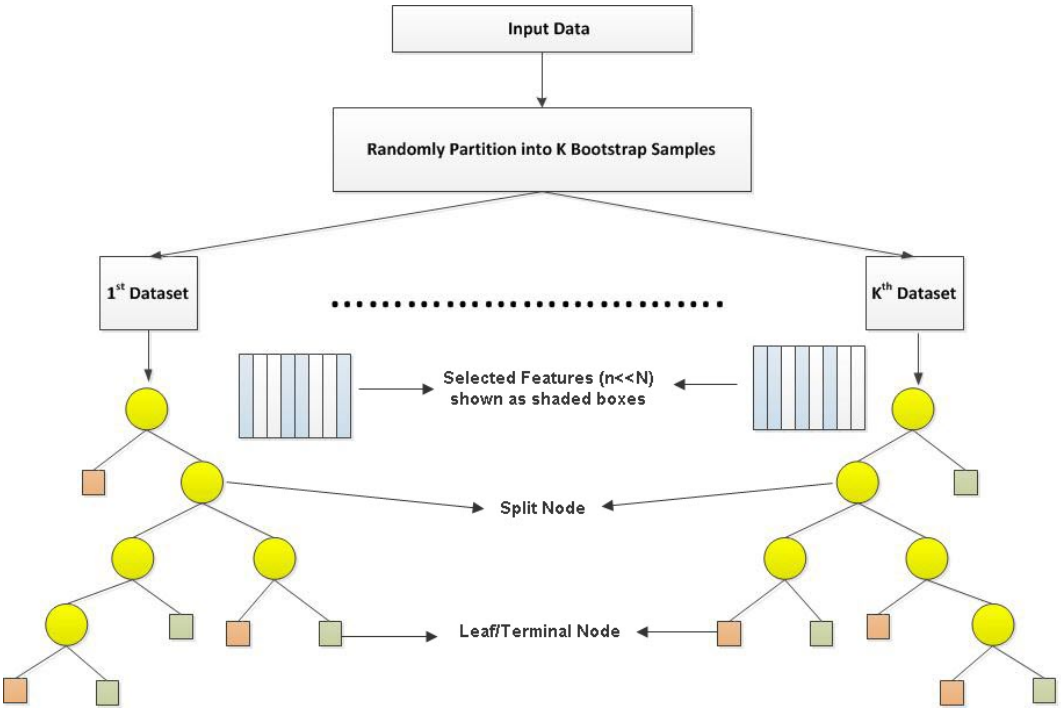


Figure 2.7: Classification trees grown by Random Forest. At each split node, shown in yellow, randomly $n \ll N$ features are selected and the node is split into two classes. Leaf nodes or terminal nodes, shown in red/green, are the nodes which cannot be split further indicating that the tree is fully grown. K such trees are grown and stored as learned model.

[8]. Given a training data $\boldsymbol{X} \in \mathbb{R}^{M \mathrm{X} N}$, the following steps are involved in growing a classification tree:

- A bootstrap sample, $\boldsymbol{X}_{\mathrm{B}} \in \mathbb{R}^{M \mathrm{X} N}$, is generated from the given training data, using sampling with replacement method, and is used as training data to grow the tree.
- A subset of $n \ll N$, features is selected at random at each node of the tree. The node is split based on the best split of these n features. The size n remains constant throughout the forest generation.
- Each tree is grown fully, without pruning.

K classification trees are grown as shown in Figure 2.7, where K can be any value; the default value used is 500. An unlabeled instance is passed down to all the decision trees in the forest and each tree assigns the new instance to one of the class-

es. This is called voting; the new instance is assigned to the class which gets majority votes. RF is one of the most accurate learning algorithms. It can efficiently handle high-dimensional (upto a couple of thousand features) and large scale datasets. It can effectively estimate missing values, balance error in unbalanced dataset, and can be extended to unlabeled data. However, unlike decision tree, the classifications made by RF are difficult to comprehend. The problem of overfitting is observed when the data is noisy. The matlab implementation of random forest algorithm is obtained from [24].

2.6 Performance Measures

This section discusses the various performance metrics used to evaluate the efficacy of the learned model. Confusion matrix for binary classification problems is shown in Table 2.1.

Table 2.1: Confusion Matrix

Test Outcome	**Condition**	
Positive	True Positive(TP)	False Positive(FP)
Negative	False Negative(FN)	True Negative(TN)

where TP refers to number of samples correctly identified as positive, FP refers to number of samples incorrectly identified as positive, TN refers to number of samples correctly identified as negative, and FN refers to number of samples incorrectly identified as negative.

False Positive Rate (FPR) is the probability of falsely rejecting null hypothesis when it is true. It is also known as Type I error and is defined as:

$$FPR = \frac{FP}{FP + TN} \tag{2.18}$$

Accuracy is defined as percentage of true results and is given by the formula:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \text{ X } 100 \tag{2.19}$$

Sensitivity, also known as recall rate or True Positive Rate (TPR), is the ratio of people

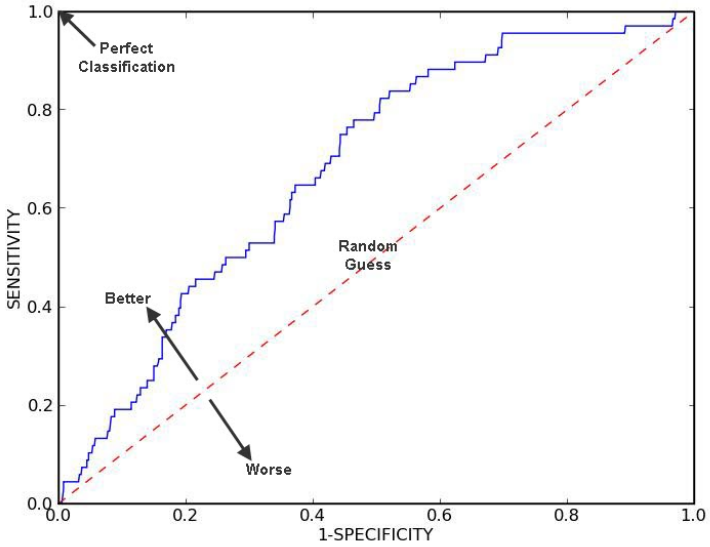


Figure 2.8: Receiver-Operating Characteristic (ROC) space. The area under the ROC curve is known as Area Under the Curve (AUC).

who are correctly identified as positive. It is computed using the formula:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.20}$$

Specificity, is the ratio of people who are correctly identified as belonging to negative class. It is know known as 1 minus False Positive Rate(FPR). It is computed using the formula:

$$Specificity = \frac{TN}{TN + FP} \tag{2.21}$$

Area under the curve (AUC) is defined as the area under Receiver-Operating Characteristic (ROC) curve. A ROC curve is created by plotting TPR as a function of FPR as illustrated in Figure 2.8. The diagonal line, also known as random guess line, defines the benchmark for classification results. Points above the diagonal are good, those on the diagonal are random guesses, and points below the diagonal are sub-optimal.

CHAPTER 3

EXPERIMENTS

This chapter gives a detailed account of experiments performed and results obtained. The following feature selection algorithms were analyzed in this study: Student′s t-test (T-test), Relief-F, FisherKernel, GiniIndex, Information Gain (InfoGain), Kruskal-Wallis, Chi-Square, Minimum Redundancy Maximum Relevance (mRMR), and Sparse Logistic Regression with Stability Selection (SSLogisticR). Random Forest (RF) and Support Vector Machines (SVM) classifiers were used to analyze the efficacy of the feature selection algorithms.

3.1 Dataset: Alzheimer′s Disease Neuroimaging Initiative (ADNI)

In the ADNI longitudinal study, blood plasma proteins obtained from an RBM panel (RBM), Magnetic Resonance Imaging (MRI), and Psychometric assessment scores (META) biomarkers[1] were studied at baseline. Table 3.1 lists the number of subjects in each disease category. NL refers to control (healthy) subjects, MCI refers to patients with Mild Cognitive Impairment, and AD are patients identified with the debilitating disease. MCI subjects are further classified as MCI-NC and MCI-C based on their disease status in the longitudinal study. MCI non-converters (MCI-NC) are those MCI subjects which remain at MCI status and MCI converters (MCI-C) are those MCI subjects who subsequently progress to AD.

Table 3.1: ADNI Baseline Data Details

Features	Feature #	NL	MCI-NC	MCI-C	AD	Total
RBM	147	58	233	163	112	566
MRI	305	191	177	142	138	648
META	52	218	220	153	171	762
RBM+MRI	452	48	176	142	83	449
RBM+META	199	55	219	153	103	530
MRI+META	357	182	166	133	124	605
RBM+MRI+META	504	45	165	133	78	421

[1] Refer Appendix A for complete list of features.

In ADNI proteomics Study, 95 plasma protein features were used at baseline. There are 54 control, 218 MCI-NC, 162 MCI-C, and 111 AD subjects. This study intends to develop an algorithm on ADNI dataset which can be successfully applied across other plasma protein based datasets such as AIBL.

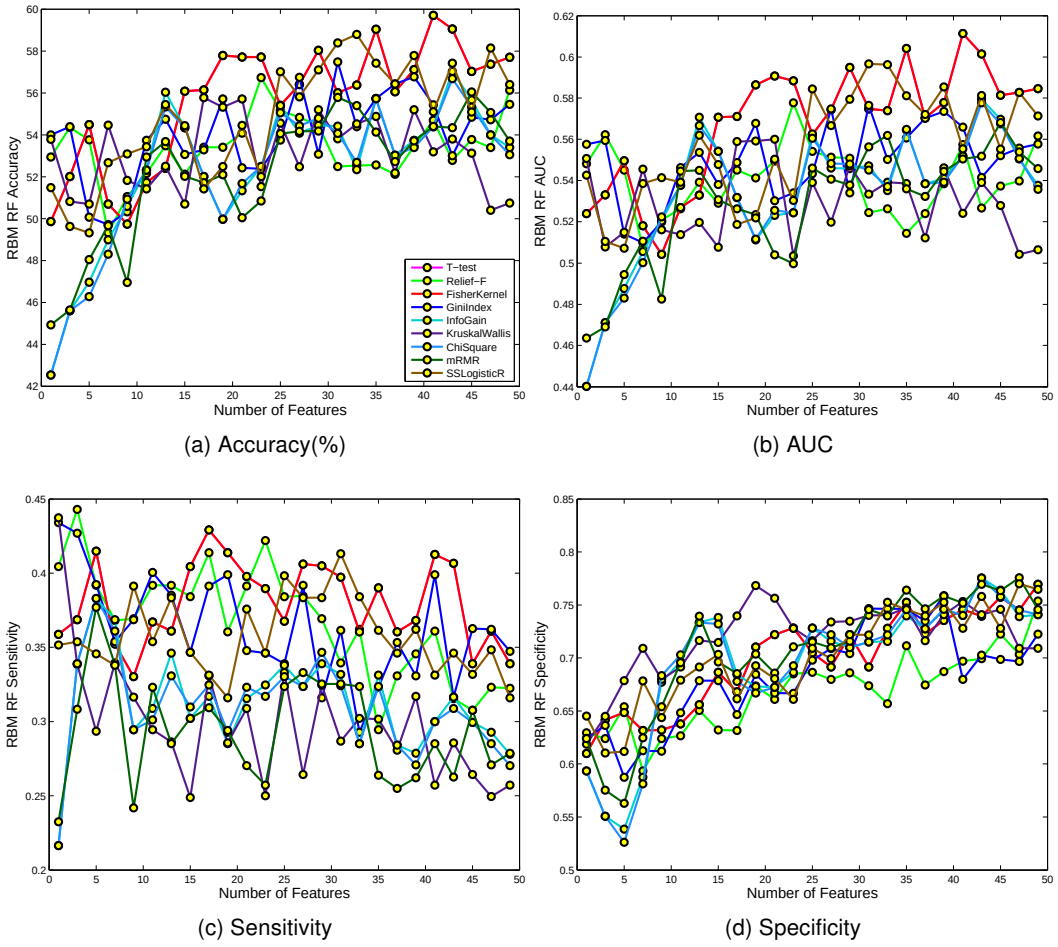
Samples with missing values were discarded from the study and the data matrix was normalized using Equation 2.2 and was partitioned into 10-datasets. Feature selection algorithms were used to rank features according to their significance. RF and SVM were used to build learning models using top 1,2,3,...,50 features for each FS algorithm. The learned model was tested on test set and classification measures were analyzed. The results reported are averaged over 10-datasets. The graphs are drawn for each performance metric as a function of number of features. The key results have been included in this chapter.

3.2 ADNI RBM, MRI, and META : MCI-NC vs MCI-C

In this experiment, the task was to identify signatures for predicting MCI-NC (negative class, 165 members) from MCI-C (positive class, 133 members) subjects. Subjects which have non-missing values for RBM, MRI, and META biomarkers and belong to either MCI-NC or MCI-C category were used. The dataset was partitioned into 10 cross-folds (subsets) such that each fold was approximately balanced with respect to the two classes. One fold was used for testing and rest of the 9 folds were used for training, creating 10 train-test cross validation sets. Area Under the Curve (AUC) was computed using True Positive Rate (TPR) and False Positive Rate (FPR) for all samples. Seven experiments were designed for RBM, MRI, META and their valid combinations, each using same set of samples (a total of 298 subjects), with different set of features.

3.2.1 Results

Protein biomarkers (RBM) gave a maximum of 59.7% classification accuracy for RF and 60.5% for SVM. T-test, Relief-F, FisherKernel, GiniIndex, and SSLogisticR performed



(a) Accuracy(%)

(b) AUC

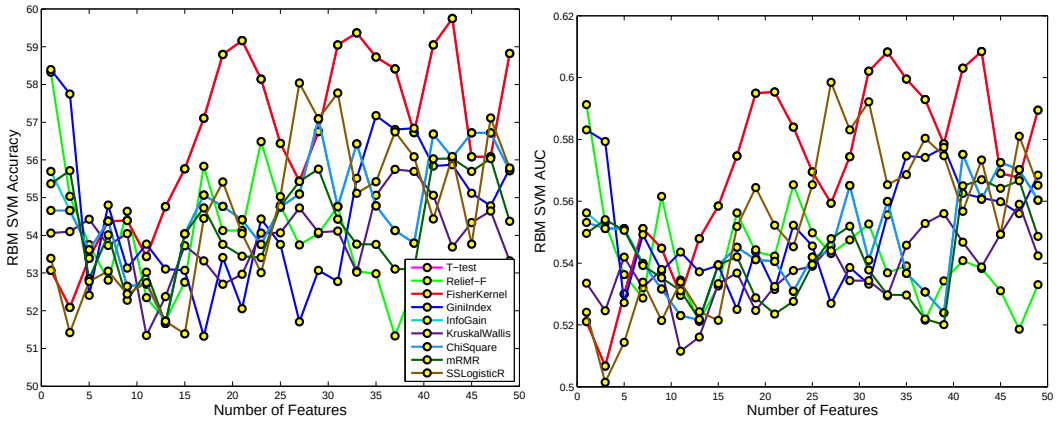
(c) Sensitivity

(d) Specificity

Figure 3.1: MCI-NC vs MCI-C Classification using RBM features & RF Classifier.

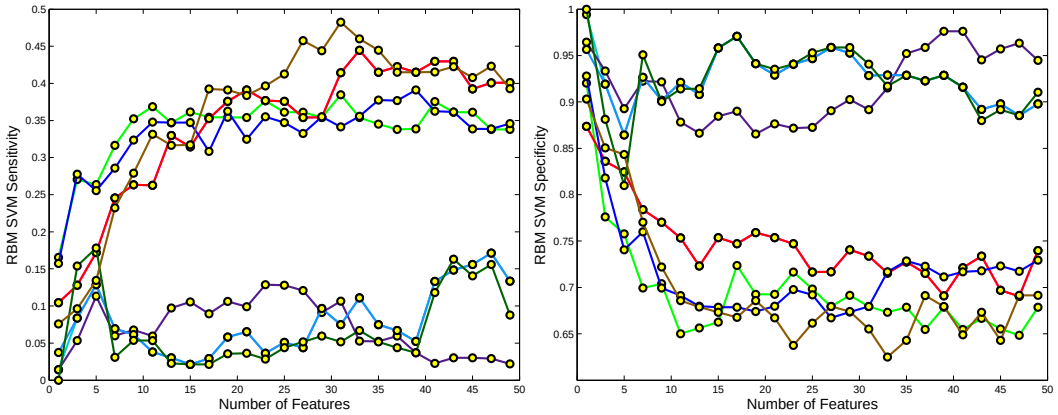
the best. Relief-F gave 59% accuracy using just one feature. As number of features were increased, its accuracy fluctuated and dropped. It gave comparatively low sensitivity and the same was observed for GiniIndex. The other three FS algorithms: T-test, FisherKernel, and SSLogisticR showed slow but steady progress with, comparatively, good AUC ($\approx$0.6). T-test and FisherKernel showed the same performance. Figures 3.1 and 3.2 show the performance measures for all algorithms. Table 3.2 lists the top 20 features selected by the top feature selection algorithms.

Figure 3.3 demonstrates that 6 out of the 10 analytes obtained by SSLogisticR were also identified by Daniel *et al.* [25] in a similar study. Many of these features have been researched directly or indirectly in context of AD. The unique features identified by



(a) Accuracy(%)

(b) AUC



(c) Sensitivity

(d) Specificity

Figure 3.2: MCI-NC vs MCI-C Classification using RBM features & SVM Classifier.

SSLogisticR in this research were extensively studied to evaluate their relevance and role in pathogenesis of AD. B Lymphocyte Chemoattractant (BLC) protein is encoded by CXCL13 gene in humans. It is linked to Neuroborreliosis, a disorder of central nervous system caused by Lyme infection, which has symptoms similar to AD. Interleukin-6 Receptor (IL_6r), a protein complex, is shown to bind with T-lymphocyte in patients with dementia similar to AD [7]. Betacellulin (BTC) has been shown to promote brain cell regeneration in mice [18]. Heparin-Binding EGF-Like Growth factor (HB-EGF) has been shown to interact with BAG1 protein produced by BAG gene which has been implicated in age-related neurodegenerative diseases.

Integrating RBM and MRI biomarkers increased the prediction accuracy of the

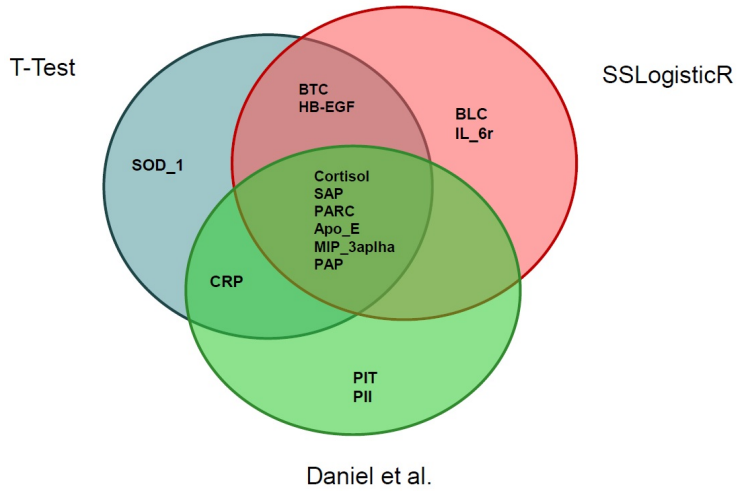
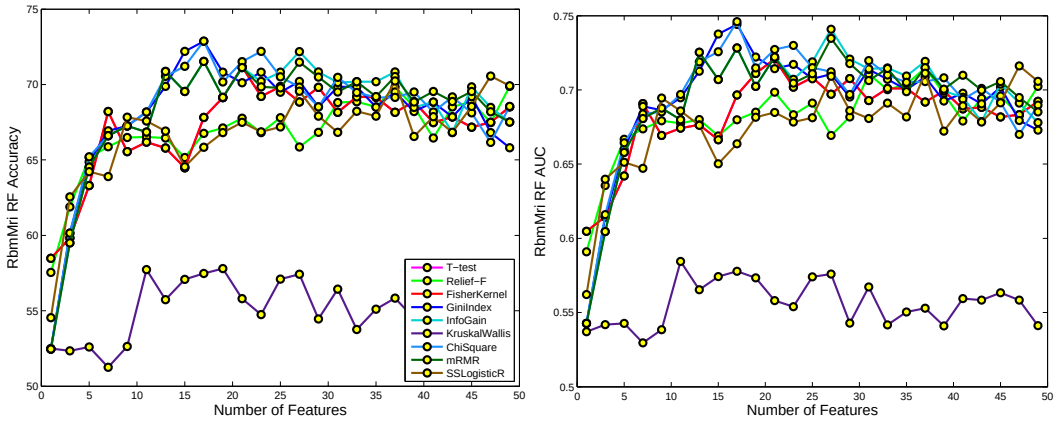


Figure 3.3: MCI-NC vs MCI-C: Comparing top 10 RBM Biomarkers identified by T-Test, SSLogisticR, and Daniel *et al.* [25]

Table 3.2: Top 20 RBM Features

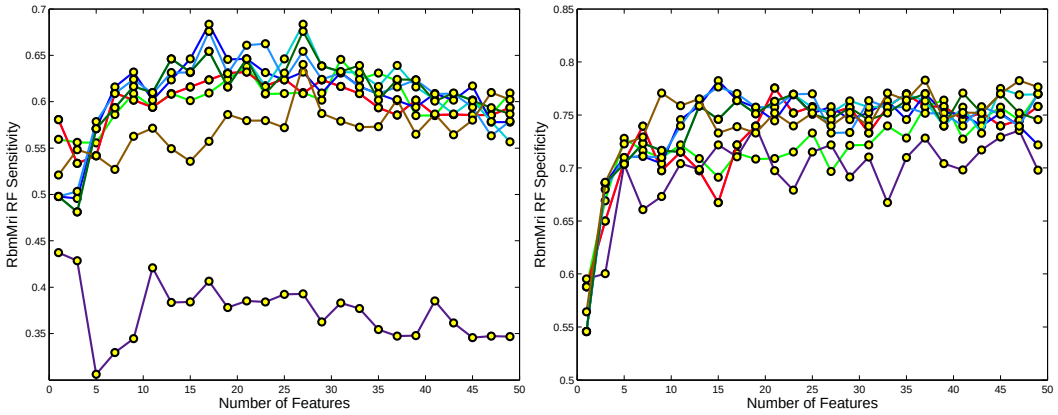
T-Test	Relief-F	GiniIndex	SSLogisticR
Apo_E	Apo_C_III	CD40_L	Cortisol
BTC	CD40_L	HBEGF	SAP
Cortisol	HBEGF	PARC	BTC
CRP	AACT	Apo_C_III	Apo_E
MIP_3alpha	Cortisol	BTC	MIP_3alpha
PAP	MIP_1alpha	FABP	PARC
SAP	EGF	IL_16	PAP
HBEGF	IL_6r	LH	BLC
PARC	ILGFBP	MIP_3alpha	HBEGF
SOD_1	PDGF_BB	Apo_A_II	IL_6r
CD40_L	PARC	AACT	FABP
IL_6r	Apo_E	ILGFBP	CRP
MMP_2	SAP	MIP_1alpha	Nr_CAM
Nr_CAM	TNFAILR	PAP	PYY
PDGF_BB	VKDPS	AAT	Apo_C_III
Apo_C_III	ANG_2	Cortisol	Apo_C_II
IL_8	CRP	CRP	TNFAILR
Apo_C_II	Insulin	Sortilin	CD40_L
FABP	TBG	ANG_2	THP
IL_16	BTC	Apo_E	TECK

algorithm by 12%, resulting in the maximum test accuracy of 72.8%. RF overshadowed SVM in this case as shown in Figures 3.4 and 3.5. With the exception of Kruskal-Wallis, all algorithms performed equally well. SSLogisticR was the only algorithm which showed smooth accuracy curve as number of features were increased. Table 3.3 lists the top 20 features selected by the top algorithms. Many studies have shown the



(a) Accuracy(%)

(b) AUC



(c) Sensitivity

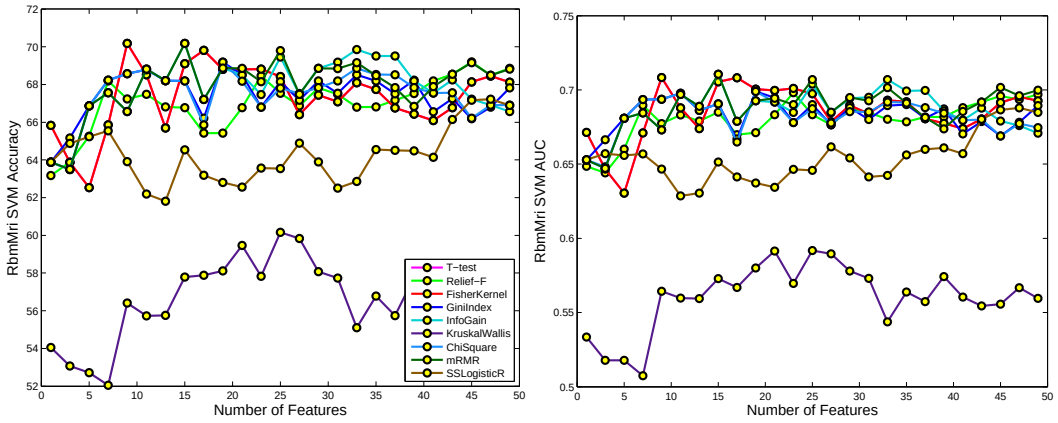
(d) Specificity

Figure 3.4: MCI-NC vs MCI-C Classification using RBM+MRI features & RF Classifier.

potential of MCI biomarkers in predicting pre-clinical AD. Hence, it was not surprising that the top features, by univariate feature selection algorithms, were dominated by MCI biomarkers. Sparse Logistic Regression (SSLogisticR) was able to capture the correlation between RBM and MCI features.

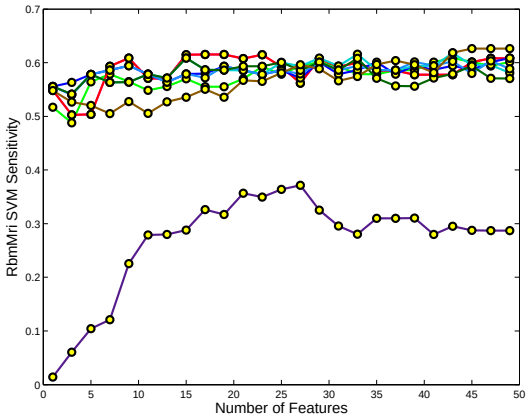
Combining RBM and META biomarkers did not show any significant improvement, reporting an accuracy increase of only 4%. Figures 3.6 and 3.7 illustrate the performance measures of different feature selection algorithms using RF and SVM respectively. MRI and META features individually gave 60% to 70% accuracy. However, no significant improvement was observed when these features were combined.

Integrating RBM, MRI, and META biomarkers resulted in a significant improve-

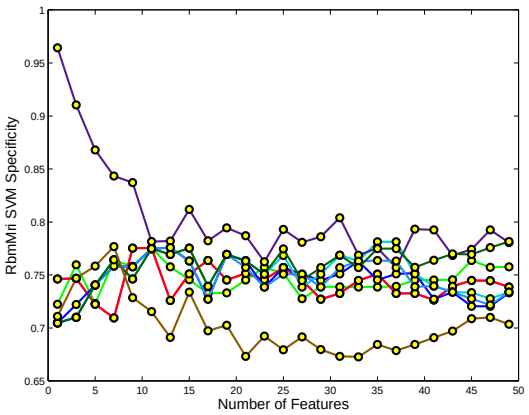


(a) Accuracy(%)

(b) AUC



(c) Sensitivity



(d) Specificity

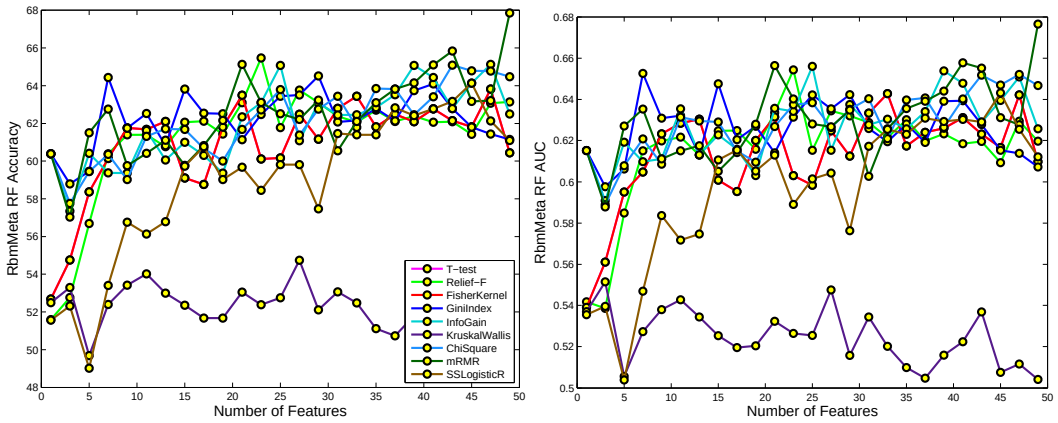
Figure 3.5: MCI-NC vs MCI-C Classification using RBM+MRI features & SVM Classifier.

ment over RBM-only accuracy (17% increase), with comparable performance from RF and SVM. Except Kruskal-Wallis, all algorithms performed equally well. The accuracy curve of most algorithms was smooth compared to previous feature combinations. This approach yielded an average AUC of 0.73, a sensitivity of 0.65 a specificity of 0.76. Table 3.4 lists the top 20 features selected by top algorithms. The features selected by SSLogisticR consisted of a balanced mixture of all three biomarkers thereby resulting in better performance. Figure 3.10 demonstrates the efficiency of integrative methods by comparing their accuracy obtained using RF for top 10 features selected by SS-LogisticR. Performances of RF and SVM for different feature selection algorithms are illustrated in Figures 3.8 and 3.9 respectively.

Table 3.3: Top 20 RBM and MRI Features

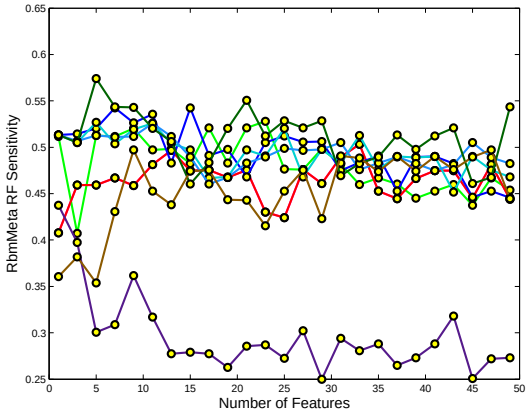
T-Test	Relief-F
Volume (WM Parcellation) of LeftAmygdala	Volume (WM Parcellation) of LeftAmygdala
Volume (Cortical Parcellation) of LeftEntorhinal	Volume (Cortical Parcellation) of LeftEntorhinal
Volume (WM Parcellation) of LeftHippocampus	Cortical Thickness Avg of LeftEntorhinal
Cortical Thickness Avg of LeftIsthmusCingulate	Cortical Thickness Avg of LeftFusiform
Cortical Thickness Avg of LeftRostralMiddle-Frontal	Volume (WM Parcellation) of LeftHippocampus
Cortical Thickness Avg of LeftTemporalPole	Cortical Thickness Avg of LeftPrecuneus
Volume (WM Parcellation) of RightAmygdala	Cortical Thickness Avg of LeftTemporalPole
Cortical Thickness Avg of RightEntorhinal	Volume (WM Parcellation) of RightAmygdala
Volume (WM Parcellation) of RightHippocampus	Cortical Thickness Avg of RightEntorhinal
Cortical Thickness Avg of RightIsthmusCingulate	Volume (WM Parcellation) of RightHippocampus
Cortical Thickness Avg of RightMiddleTemporal	Cortical Thickness Avg of RightMiddleTemporal
Cortical Thickness Avg of RightInsula	Cortical Thickness Avg of LeftIsthmusCingulate
Volume (Cortical Parcellation) of RightInferior-Parietal	Cortical Thickness Avg of LeftMiddleTemporal
Cortical Thickness Avg of RightPrecuneus	Cortical Thickness Avg of RightInferiorParietal
Cortical Thickness STDV of RightPrecuneus	Cortical Thickness Avg of RightIsthmusCingulate
Cortical Thickness Avg of LeftEntorhinal	Volume (Cortical Parcellation) of RightMid-dleTemporal
Cortical Thickness Avg of LeftMiddleTemporal	Cortical Thickness Avg of LeftInferiorParietal
Cortical Thickness Avg of LeftInsula	Cortical Thickness Avg of LeftInferiorTemporal
Cortical Thickness Avg of RightInferiorParietal	Cortical Thickness Avg of LeftRostralMiddle-Frontal
Cortical Thickness Avg of RightInferiorTemporal	Volume (Cortical Parcellation) of LeftInferiorTem-poral

GiniIndex	SSLogisticR
Volume (WM Parcellation) of LeftAmygdala	Volume (WM Parcellation) of LeftHippocampus
Volume (Cortical Parcellation) of LeftEntorhinal	Volume (Cortical Parcellation) of LeftEntorhinal
Cortical Thickness Avg of LeftEntorhinal	Cortical Thickness STDV of RightPrecuneus
Cortical Thickness Avg of LeftFusiform	PARC
Volume (WM Parcellation) of LeftHippocampus	Cortisol
Cortical Thickness Avg of LeftIsthmusCingulate	PYY
Cortical Thickness Avg of LeftPericalcarine	HBEGF
Cortical Thickness Avg of LeftTemporalPole	Cortical Thickness STDV of LeftBankssts
Volume (WM Parcellation) of RightAmygdala	Volume (Cortical Parcellation) of LeftTempo-ralPole
Cortical Thickness Avg of RightEntorhinal	Cortical Thickness STDV of LeftEntorhinal
Volume (WM Parcellation) of RightHippocampus	Cortical Thickness STDV of LeftPostcentral
Cortical Thickness Avg of RightMiddleTemporal	Cortical Thickness Avg of LeftIsthmusCingulate
Volume (Cortical Parcellation) of LeftInferiorTem-poral	CRP
Cortical Thickness Avg of LeftPrecuneus	PLGF
Volume (Cortical Parcellation) of LeftInferiorPari-etal	IL_16
Volume (Cortical Parcellation) of RightMid-dleTemporal	Insulin
Cortical Thickness Avg of RightSuperiorTemporal	Nr_CAM
Cortical Thickness Avg of RightIsthmusCingulate	Cortical Thickness Avg of RightEntorhinal
Cortical Thickness STDV of RightPrecuneus	MIP_3alpha
Cortical Thickness Avg of RightInsula	IL_6r

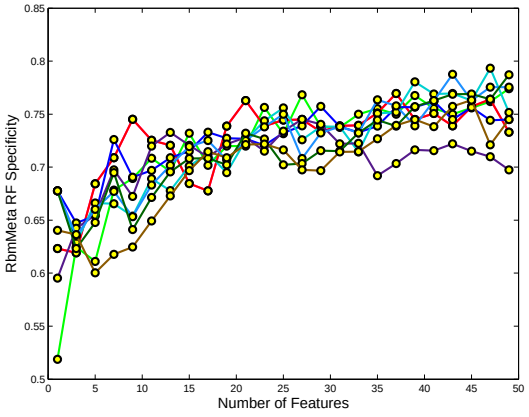


(a) Accuracy(%)

(b) AUC

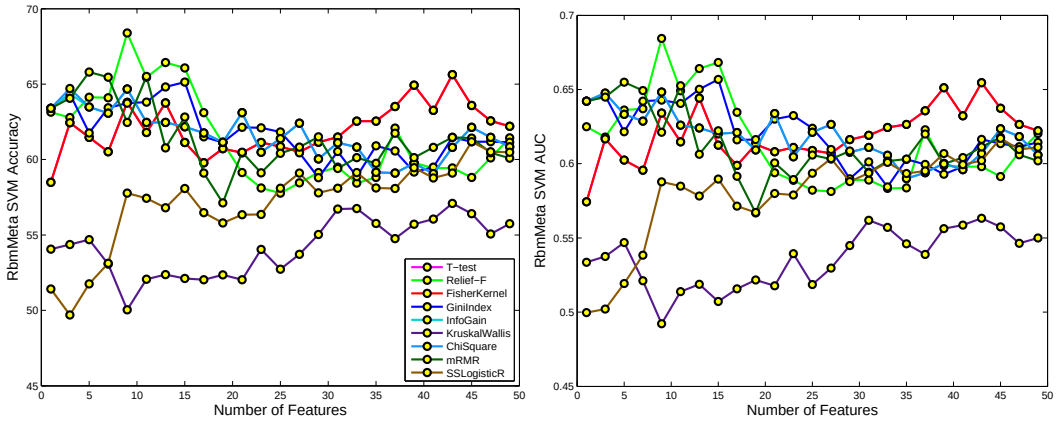


(c) Sensitivity



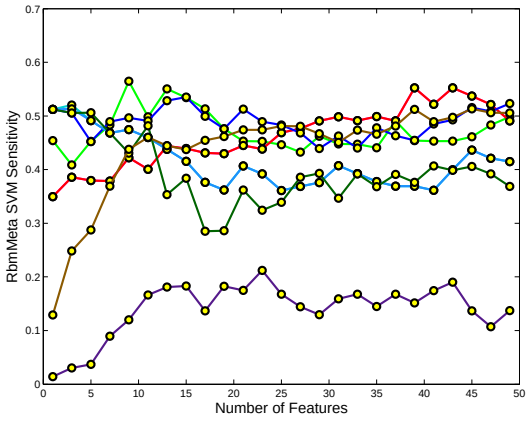
(d) Specificity

Figure 3.6: MCI-NC vs MCI-C Classification using RBM+META features & RF Classifier.

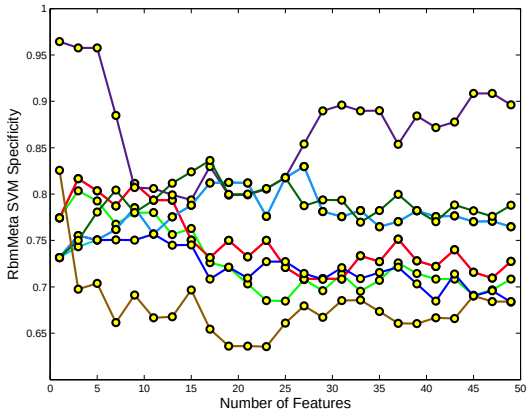


(a) Accuracy(%)

(b) AUC



(c) Sensitivity

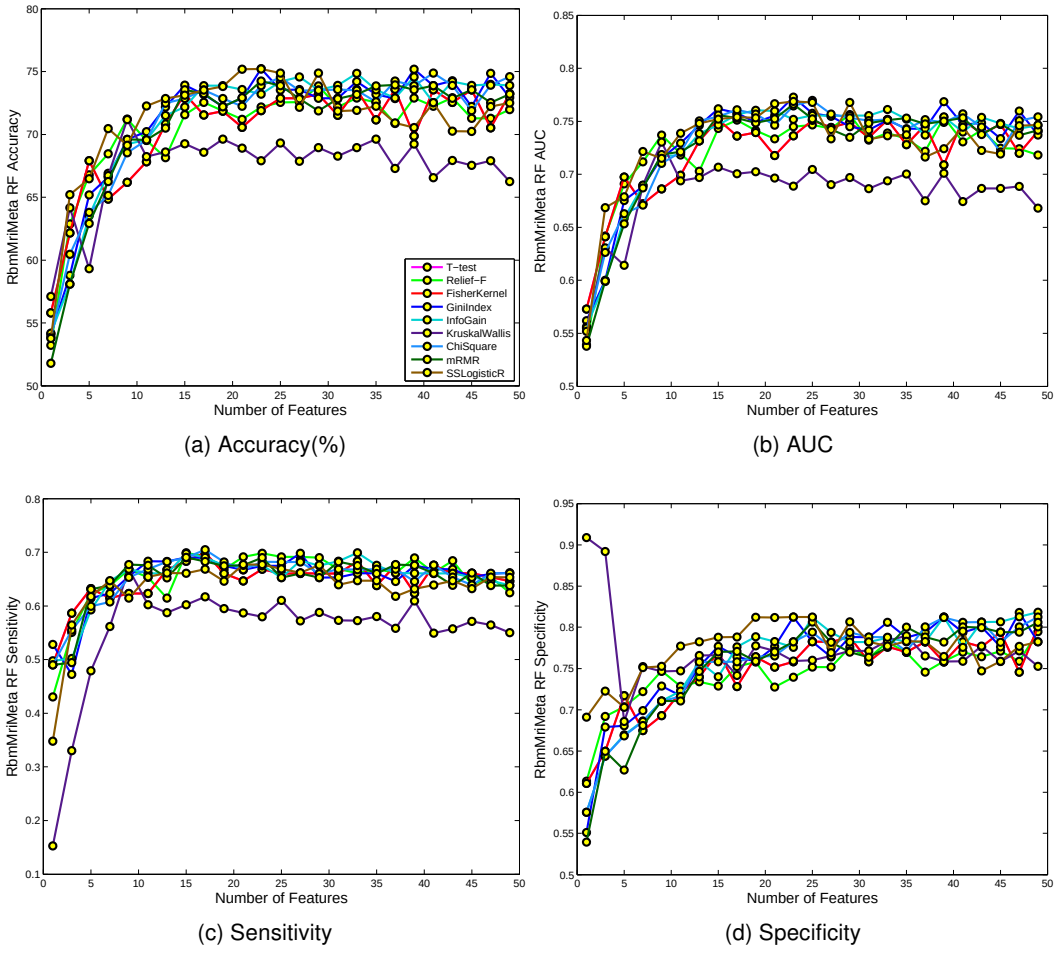


(d) Specificity

Figure 3.7: MCI-NC vs MCI-C Classification using RBM+META features & SVM Classifier.

Table 3.4: Top 20 RBM, MRI,& META Features

T-Test	**Relief-F**
Volume (WM Parcellation) of LeftAmygdala	Volume (Cortical Parcellation) of LeftEntorhinal
Volume (Cortical Parcellation) of LeftEntorhinal	Volume (WM Parcellation) of LeftHippocampus
Volume (WM Parcellation) of LeftHippocampus	Cortical Thickness Avg of LeftTemporalPole
Cortical Thickness Avg of RightEntorhinal	Volume (WM Parcellation) of RightAmygdala
Volume (WM Parcellation) of RightHippocampus	Cortical Thickness Avg of RightEntorhinal
Cortical Thickness Avg of RightMiddleTemporal	Volume (WM Parcellation) of RightHippocampus
LDELTOTAL	Cortical Thickness Avg of RightMiddleTemporal
ADAS_sub1	LDELTOTAL
ADAS_sub4	ADAS_sub1
ADAS_sub7	ADAS_sub4
FAQ	ADAS_sub7
Cortical Thickness Avg of LeftIsthmusCingulate	FAQ
Cortical Thickness Avg of LeftRostralMiddleFrontal	APOE
Volume (WM Parcellation) of RightAmygdala	Volume (WM Parcellation) of LeftAmygdala
Volume (Cortical Parcellation) of RightInferiorParietal	Cortical Thickness Avg of LeftEntorhinal
Cortical Thickness Avg of LeftTemporalPole	Cortical Thickness Avg of LeftFusiform
Cortical Thickness Avg of RightIsthmusCingulate	Cortical Thickness Avg of LeftPrecuneus
Cortical Thickness Avg of RightInsula	Cortical Thickness Avg of LeftIsthmusCingulate
Cortical Thickness Avg of LeftEntorhinal	Cortical Thickness Avg of RightIsthmusCingulate
Cortical Thickness Avg of RightInferiorTemporal	CDR
GiniIndex	**SSLogisticR**
Volume (WM Parcellation) of LeftAmygdala	FAQ
Volume (Cortical Parcellation) of LeftEntorhinal	ADAS_sub4
Cortical Thickness Avg of LeftEntorhinal	LDELTOTAL
Cortical Thickness Avg of LeftFusiform	ADAS_sub1
Volume (WM Parcellation) of LeftHippocampus	ADAS_sub7
Cortical Thickness Avg of LeftPericalcarine	CRP
Cortical Thickness Avg of LeftTemporalPole	Cortical Thickness STDV of RightPrecuneus
Volume (WM Parcellation) of RightAmygdala	Volume (Cortical Parcellation) of LeftTemporalPole
Cortical Thickness Avg of RightEntorhinal	PARC
Volume (WM Parcellation) of RightHippocampus	Volume (WM Parcellation) of LeftHippocampus
Cortical Thickness Avg of RightMiddleTemporal	Volume (Cortical Parcellation) of LeftEntorhinal
LDELTOTAL	Volume (WM Parcellation) of RightCerebellumCortex
ADAS_sub1	Cortical Thickness STDV of LeftCuneus
ADAS_sub4	PYY
FAQ	Nr_CAM
Volume (Cortical Parcellation) of LeftInferiorParietal	SAP
Volume (Cortical Parcellation) of RightMiddleTemporal	IL_16
Volume (Cortical Parcellation) of LeftInferiorTemporal	APOE
Cortical Thickness Avg of LeftPrecuneus	NPI
Cortical Thickness Avg of LeftIsthmusCingulate	Cortisol



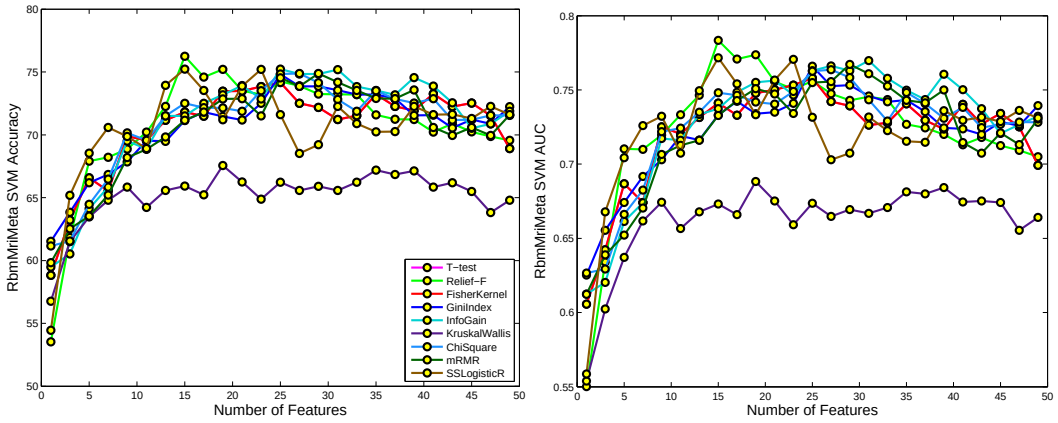
(a) Accuracy(%)

(b) AUC

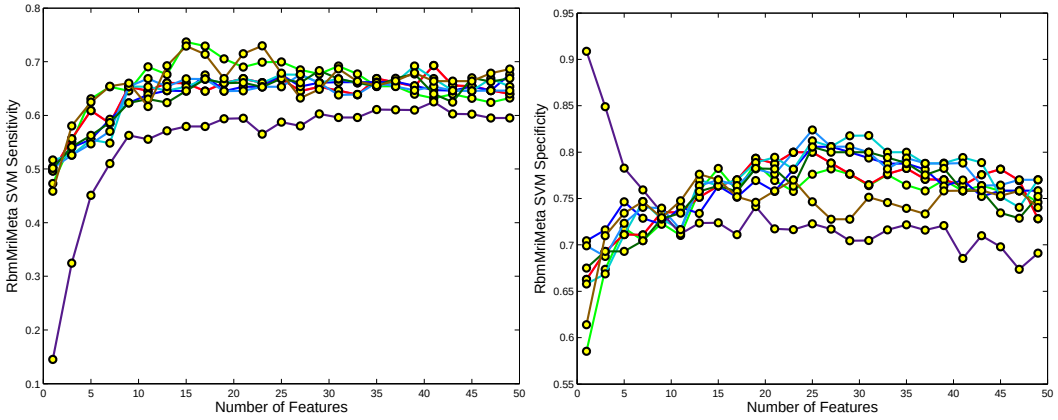
(c) Sensitivity

(d) Specificity

Figure 3.8: MCI-NC vs MCI-C Classification using RBM+MRI+META features & RF Classifier.



(a) Accuracy(%) (b) AUC



(c) Sensitivity (d) Specificity

Figure 3.9: MCI-NC vs MCI-C Classification using RBM+MRI+META features & SVM Classifier.

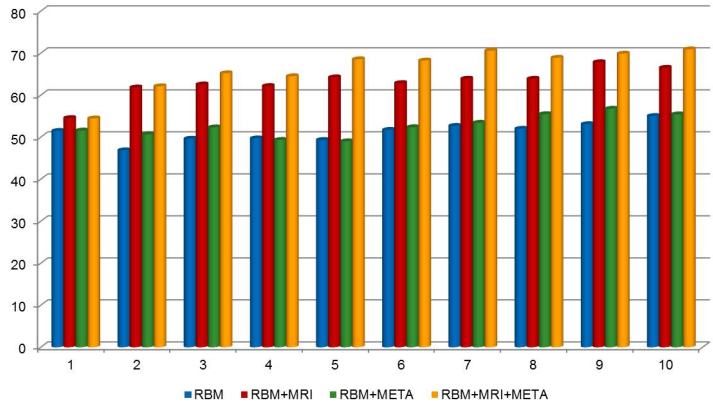


Figure 3.10: MCI-NC vs MCI-C: Comparison of RF Accuracy obtained for top 10 biomarkers selected by SSLogisticR using various Integrative Approaches.

3.3 ADNI Proteomics: NL vs AD

This is an NL vs AD classification experiment with 54 NL (negative class) and 111 AD (positive class) subjects. Since this is an unbalanced data, over and under sampling approaches are evaluated. The given unbalanced dataset is randomly partitioned into train-test set in 9:1 ratio. For each sampling method, 10 such datasets are generated. Table 3.5 gives the number of training and testing samples used in various sampling methods.

Table 3.5: NL vs AD : Dataset Details

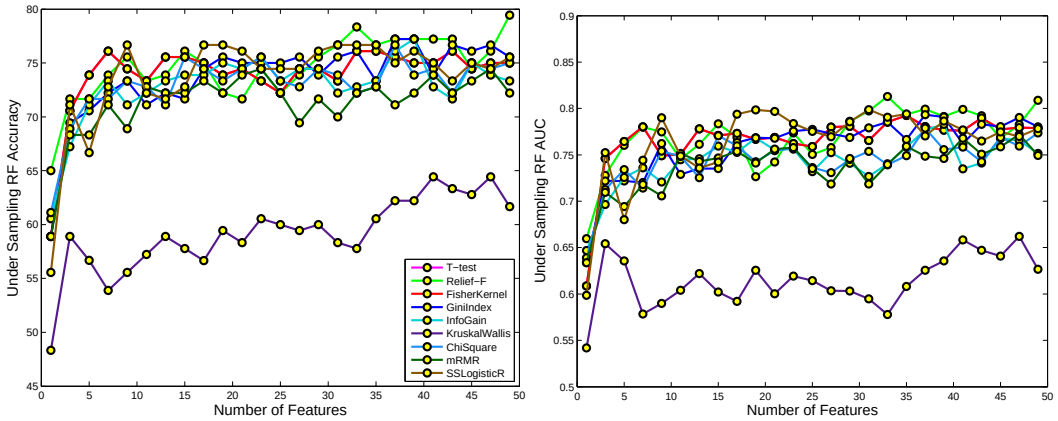
		No	Sampling	Under	Sampling	Over	Sampling
Target	**Sample #**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
AD(positive)	111	99	12	48	12	99	12
NL(negative)	54	48	6	48	6	96	6
Total	**165**	**147**	**18**	**96**	**18**	**195**	**18**

3.3.1 Results

Table 3.6: Top 20 features for NL vs AD

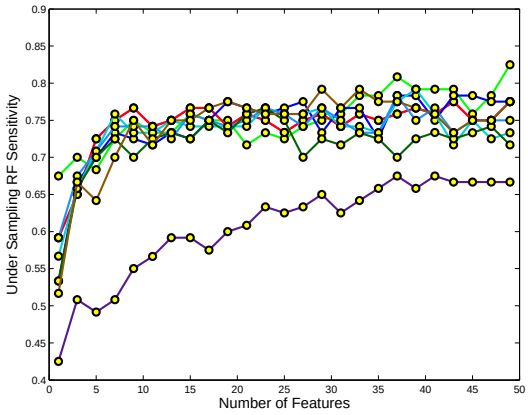
T-test	Relief-F	GiniIndex	SSLogisticR
log10(ApoE)	log10(ApoE)	log10(ApoE)	log10(ApoE)
log10(IGM)	log10(IGM)	log10(BTC)	log10(SGOT)
log10(PAPPA)	log10(MIP1alpha)	log10(PAPPA)	log10(IL16)
log10(SGOT)	log10(PAPPA)	log10(SGOT)	log10(IGM)
log10(HBEGF)	log10(SGOT)	log10(IGM)	log10(A2Micro)
log10(TIMMP1)	log10(TNC)	log10(MIP1alpha)	log10(PAPPA)
log10(TNC)	log10(A2Micro)	log10(A2Micro)	log10(ApoB)
log10(A2Micro)	log10(TIMMP1)	log10(C3)	log10(TNC)
log10(C3)	log10(C3)	log10(HGF)	log10(HBEGF)
log10(B2M)	log10(BTC)	log10(PPP)	log10(MIP1alpha)
log10(IL16)	log10(HBEGF)	log10(TIMMP1)	log10(PPP)
log10(MIP1alpha)	log10(IL16)	log10(TNC)	log10(CEA)
log10(VCAM)	log10(PPP)	log10(HBEGF)	log10(C3)
log10(PPP)	log10(VCAM)	log10(IL16)	log10(ATEN)
log10(VEGF)	log10(ACE)	log10(ApoH)	log10(ADP)
log10(Apo_C3)	log10(Apo_C3)	log10(Apo_C3)	log10(CORTISOL)
log10(ApoB)	log10(ApoB)	log10(CRP)	log10(CRP)
log10(ApoH)	log10(VEGF)	log10(ACE)	log10(VEGF)
log10(BTC)	log10(ADP)	log10(ApoB)	log10(HGF)
log10(TBG)	log10(ATEN)	log10(CORTISOL)	log10(BTC)

Results for no sampling and over-sampling approaches were comparable. Over-sampling performed better in few cases. High sensitivity (SN) of 0.95 was observed in both of these sampling approaches. Specificity (SP) was observed between 0.61 and

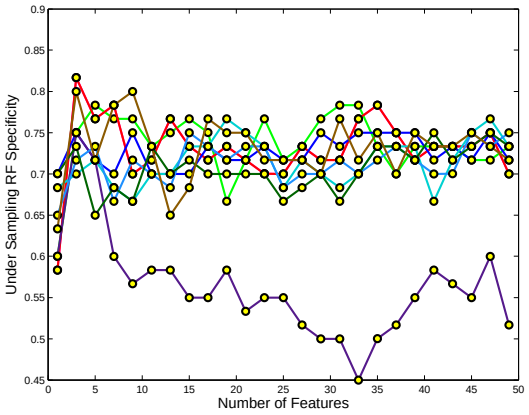


(a) Accuracy(%)

(b) AUC



(c) Sensitivity

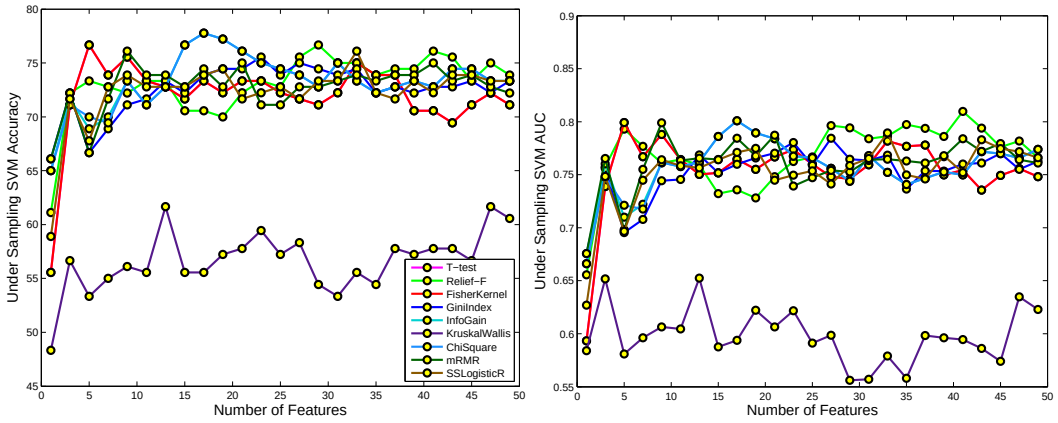


(d) Specificity

Figure 3.11: NL vs AD Classification using Under-Sampling features & RF Classifier.

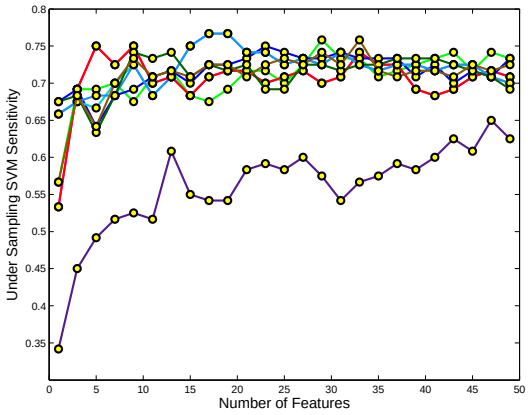
0.67 while AUC was in the range of 0.75. Test Accuracy (TA) was between 72% and 83%, but standard deviation (STDV) was very high. In light of these factors, these t-wo techniques did not give satisfactory results. SSLogisticR showed good results in the case of over-sampling using RF (TA-84.44%, STDV-6.83, AUC-0.82, SN-0.94, SP-0.65). RF performed better than SVM.

Under-sampling accuracy was $\approx$79.8%, showing high standard deviation, but comparable sensitivity and specificity (less than 0.75). Average AUC observed was 0.78. Reilef-F showed a smooth increase in performance with increasing number of features, although performance fluctuations were seen. GiniIndex did not perform well for very less number of features, but slowly gained 80% accuracy for top 45 features.

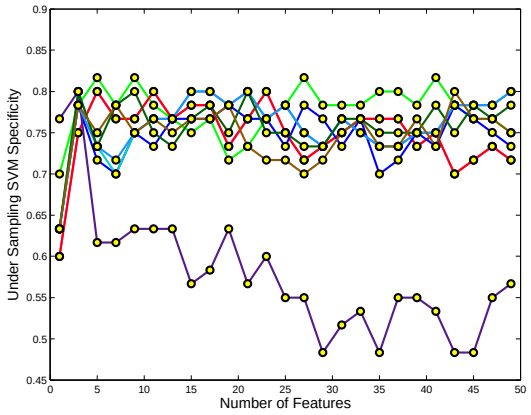


(a) Accuracy(%)

(b) AUC



(c) Sensitivity



(d) Specificity

Figure 3.12: NL vs AD Classification using Under-Sampling features & SVM Classifier.

Other algorithms fluctuated a lot and showed average performance. Performance measures of various feature selection algorithms using RF and SVM classifiers are shown in Figures 3.11 and 3.12 respectively. Table 3.6 lists the top 20 features from the top algorithms.

Daniel *et al.* [25] used all RBM features shown in table 3.1 and 54 samples from each class to find protein signatures for predicting AD subjects. They obtained 21-protein signature; 11 of these were studied in this experiment. A total of 9 analytes matched the biomarkers obtained by SSLogisticR as shown in figure 3.13. The relevance of features not previously identified was analyzed in detail. Several researchers have identified significant difference in Cortisol levels, a stress related hormone, in

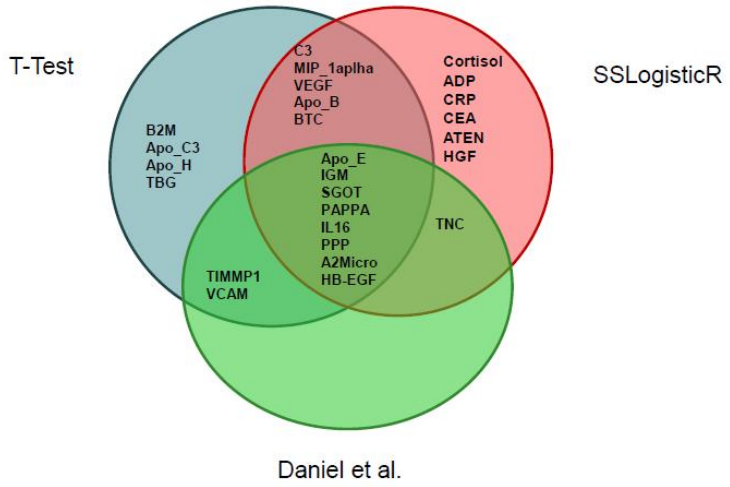


Figure 3.13: NL vs AD: Comparing top 10 RBM Biomarkers obtained using Undersampling approach by T-Test, SSLogisticR, and Daniel *et al.* [25]

healthy, MCI-NC, MCI-C, and AD patients [33][5]. Adiponectin (ADP) is a protein hormone found in body fat and has been linked to dementia [47]. Une *et al.* [46] studied the ADP levels in plasma and CSF in NL, MCI, and AD subjects. C Reactive Protein (CRP) levels rise in response to inflammation. Wood *et al.* [48] found increased levels of CRP in Alzheimer′s cortex. Macrophage inflammatory Protein 1 alpha (MIP_1alpha or CCL3) plays a crucial role in immune response and has a proinflammatory effect. Tripathy *et al.* [43] found increased levels of MIP_1alpha in Alzheimer′s Vessels. Carcinoembryonic antigen (CEA) is a glycoprotein involved in cell adhesion and mainly studied in context of cancerous tumors. Licastro *et al.* [30] studied the role of environmental factors in pathogenesis of AD and found CEA to be a key modulator in neuronal loss, inflammation, and amyloid deposition. Angiotensinogen (ATEN or AGT), also known as renin substrate, is an α-2 globulin and its role in neurodegenerative disorders has been studied by Savaskan *et al.* [41]. Hepatocyte growth factor (HGF) regulates cell growth, cell motility, and morphogenesis. Tsuboi *et al.* [44] found that white matter damage in AD brain is correlated to CSF HGF levels. Tenascin C (TNC) has anti cell adhesive properties and its role in neuroprotection has been extensively studied by Alzheimer and Werner [4]. Apolipoprotein B (Apo_B) has been studied in context of heart disease. Caramelli *et al.* [10] noted high serum concentration of Apo_B in AD subjects.

3.4 ADNI Proteomics: NL vs MCI

In this experiment, both MCI-NC and MCI-C samples were used in positive class, NL being the negative class. The dataset setup was similar to Section 3.3. Table 3.7 gives the break-up of training and testing data used in various sampling methods:

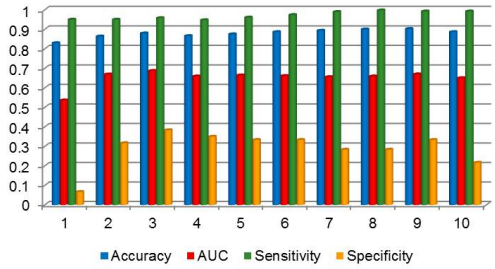
Table 3.7: NL vs MCI : Dataset Details

		No	**Sampling**	**Under**	**Sampling**	**Over**	**Sampling**
Target	**Sample #**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
MCI_All(positive)	380	342	38	48	38	342	38
NL(negative)	54	48	6	48	6	336	6
Total	**165**	**147**	**18**	**96**	**18**	**195**	**18**

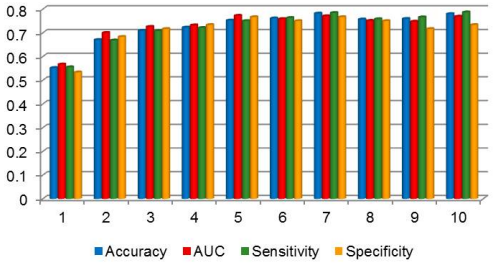
3.4.1 Results

When no sampling methods were used, test accuracy was as high as 95%. AUC was very low and results showed very high sensitivity, close to 0.98, and low specificity indicative of the dominance of positive samples. None of the feature selection algorithms were able to find good set of features in this case. SSLogisticR proved to be more stable than others. Over-sampling method initially looked promising, but after 3 to 5 features it started depicting high sensitivity and low specificity. AUC statistics were better in over-sampling than in no sampling.

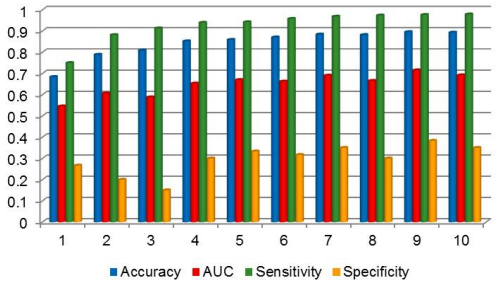
Under-sampling gave the best results. Almost all algorithms showed greater than 74% accuracy using the top four biomarkers. The overall maximum test accuracy was 83%; other performance statistics were also good. SSLogisticR was the best performing algorithm giving the highest accuracy of 84.5%. The algorithms ordered by their performance are as follows: SSLogisticR, GiniIndex, T-Test, Relief-F, FisherKernel, InfoGain, ChiSquare, KruskalWallis, mRMR. InfoGain was better than GiniIndex initially, but it fluctuated and showed no increase in performance as the number of features were increased. RF performed better than SVM. Table 3.8 shows the top 20 features selected by the top algorithms for under-sampling. Performance measures of different feature selection algorithms using RF and SVM classifiers are illustrated in Figures 3.15 and 3.16 respectively. Comparison of different sampling approaches is illustrated in Figure



(a) No Sampling



(b) Under Sampling

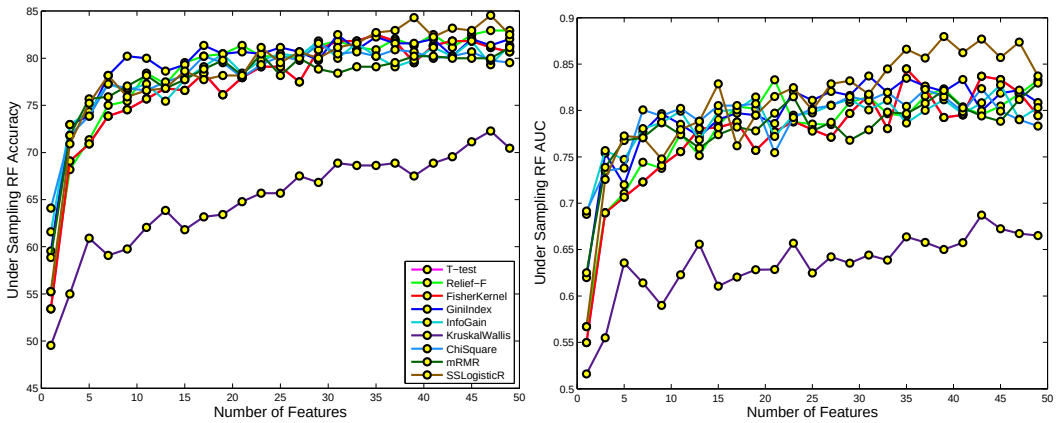


(c) Over Sampling

Figure 3.14: NL vs MCI: Comparison of different sampling approaches using RF performance measures for top 10 features identified by SSLogisticR

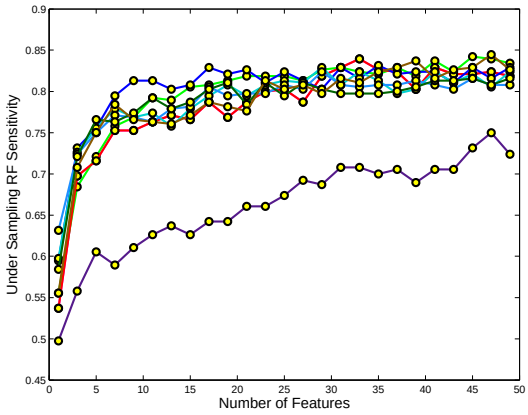
Table 3.8: Top 20 Features for NL vs MCI

T-test	Relief-F	GiniIndex	SSLogisticR
log10(ApoE)	log10(ApoE)	log10(ApoE)	log10(ApoE)
log10(MIP1alpha)	log10(HBEGF)	log10(HBEGF)	log10(MIP1alpha)
log10(PAPPA)	log10(MIP1alpha)	log10(MIP1alpha)	log10(ApoD)
log10(ANG2)	log10(ANG2)	log10(ANG2)	log10(PARC)
log10(ApoD)	log10(FASL)	log10(BTC)	log10(PAPPA)
log10(FASL)	log10(PAPPA)	log10(SGOT)	log10(ANG2)
log10(FAC7)	log10(A2Micro)	log10(A2Micro)	log10(HBEGF)
log10(HBEGF)	log10(ACE)	log10(FASL)	log10(ACE)
log10(Leptin)	log10(IL16)	log10(IL16)	log10(ATEN)
log10(PARC)	log10(Leptin)	log10(PPP)	log10(FASL)
log10(ACE)	log10(SGOT)	log10(CRP)	log10(IL16)
log10(ATEN)	log10(ApoD)	log10(HGF)	log10(A2Micro)
log10(IL16)	log10(BTC)	log10(ACE)	log10(CgA)
log10(PPP)	log10(CRP)	log10(ApoD)	log10(PPP)
log10(CRP)	log10(C3)	log10(Leptin)	log10(TEST)
log10(IGM)	log10(IGM)	log10(PAPPA)	log10(NrCAM)
log10(IIGFBP)	log10(PPP)	log10(TNC)	log10(CRP)
log10(BMP_6)	log10(Apo_C3)	log10(ATEN)	log10(Leptin)
log10(CgA)	log10(ATEN)	log10(C3)	log10(BTC)
log10(FRTN)	log10(BMP_6)	log10(FAC7)	log10(RESISTIN)

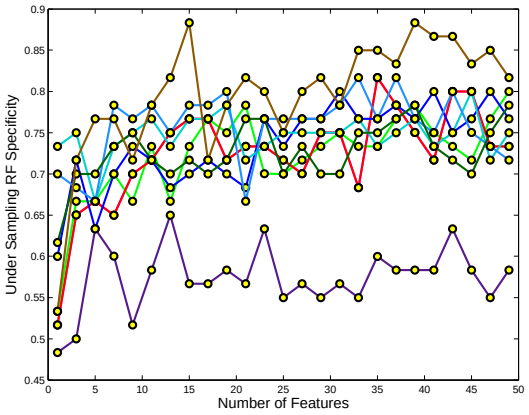


(a) Accuracy(%)

(b) AUC



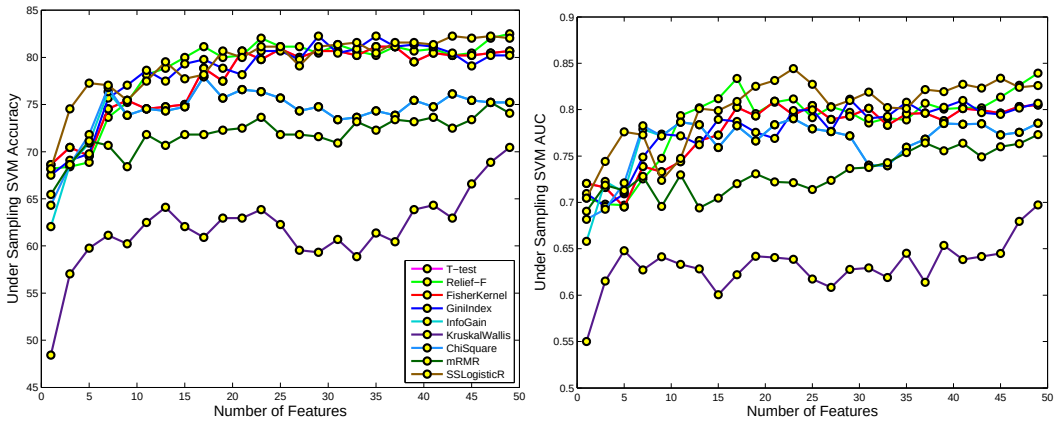
(c) Sensitivity



(d) Specificity

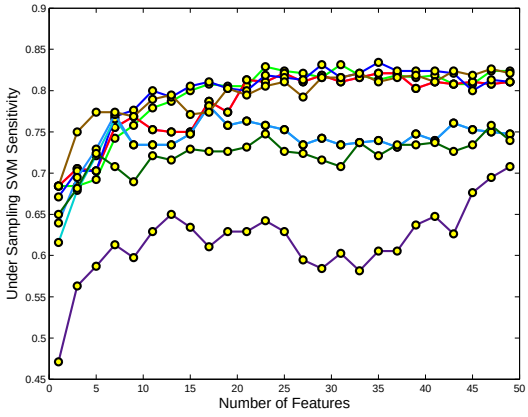
Figure 3.15: NL vs MCI Classification using Under-Sampling features & RF Classifier.

3.14 for top 10 features obtained by SSLogisticR and classified by RF. Results obtained using no sampling and over-sampling approaches depict high accuracy and sensitivity but poor specificity. In comparison, under-sampling technique resulted in reasonable accuracy in conjunction with good sensitivity and specificity. Similar results have also been noted in a recent work by Chawla *et al.* [12].

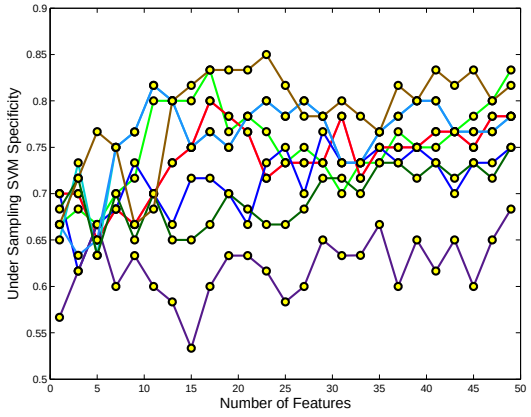


(a) Accuracy(%)

(b) AUC



(c) Sensitivity



(d) Specificity

Figure 3.16: NL vs MCI Classification using Under-Sampling features & SVM Classifier.

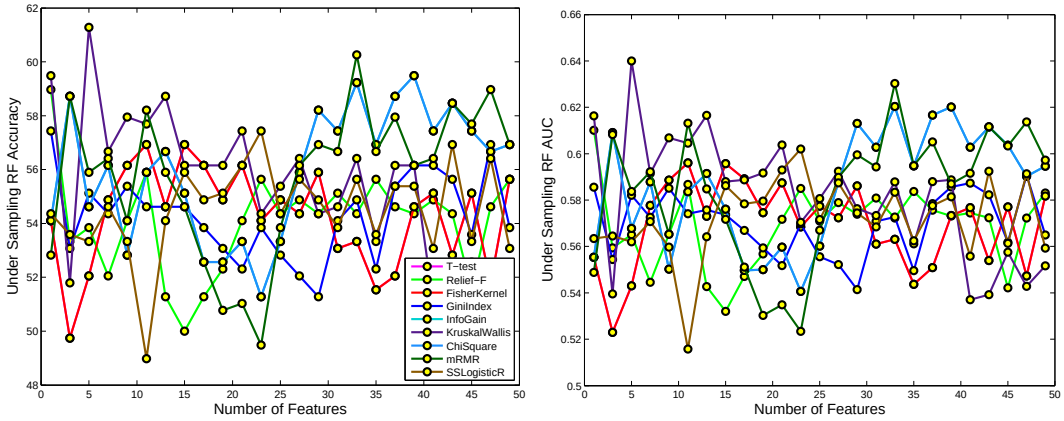
3.5 ADNI Proteomics: MCI-NC vs MCI-C

In this experiment, MCI-NC samples belong to negative class and MCI-C samples were used as positive class. The dataset setup is similar to Section 3.3. Table 3.9 gives the break-up of training and testing data used in various sampling methods.

Table 3.9: MCI-NC vs MCI-C : Dataset Details

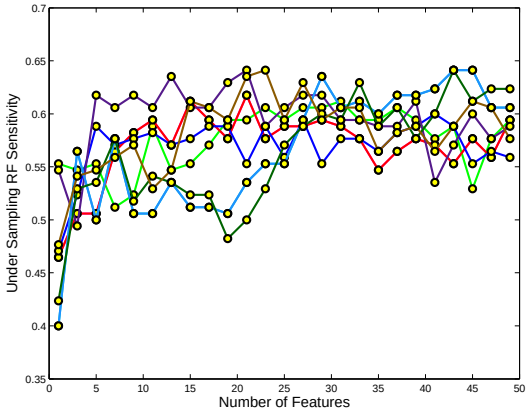
		No	Sampling	Under	Sampling	Over	Sampling
Target	**Sample #**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
MCI-C(positive)	162	145	17	145	17	195	17
MCI-NC(negative)	218	196	22	145	22	196	22
Total	**165**	**147**	**18**	**96**	**18**	**195**	**18**

3.5.1 Results

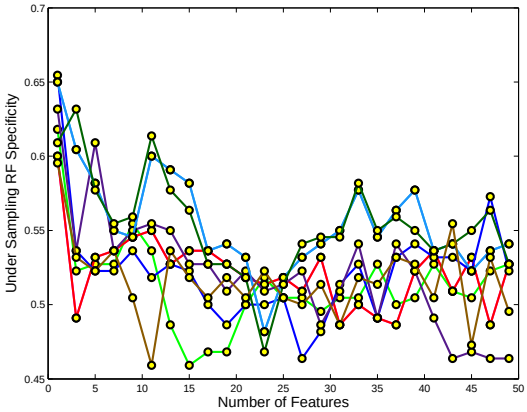


(a) Accuracy(%)

(b) AUC

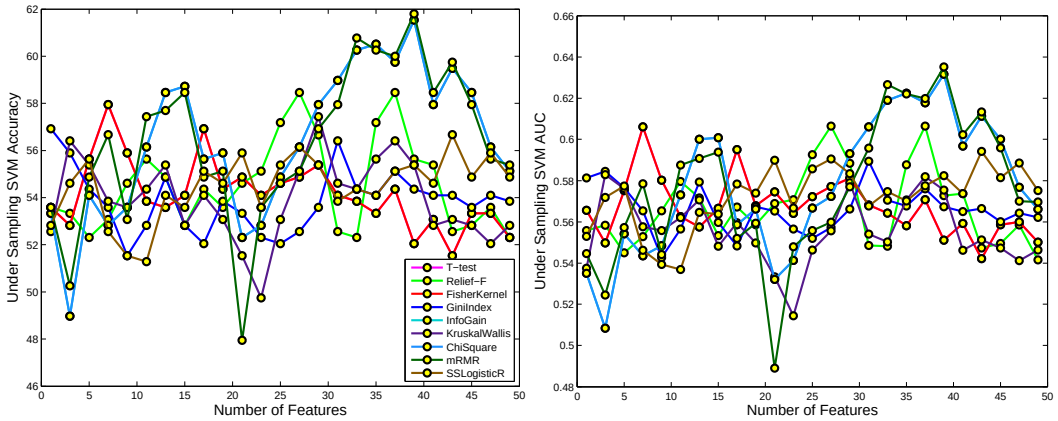


(c) Sensitivity



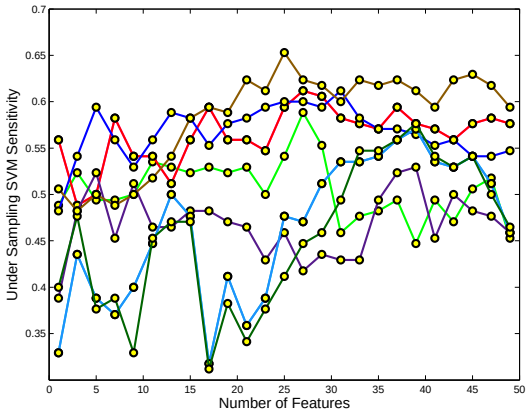
(d) Specificity

Figure 3.17: MCI-NC vs MCI-C Classification using Under-Sampling features & RF Classifier.

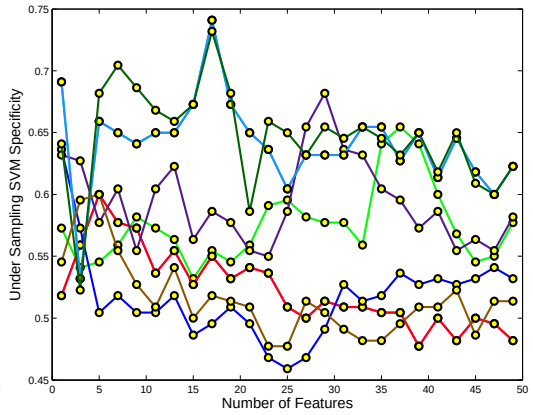


(a) Accuracy(%)

(b) AUC



(c) Sensitivity



(d) Specificity

Figure 3.18: MCI-NC vs MCI-C Classification using Under-Sampling features & SVM Classifier.

Predicting MCI-NC from MCI-C subjects is a difficult task due to heterogeneity of MCI-NC category. Effect of large number of negative samples is clearly evident from the high specificity of 0.82 and low sensitivity of 0.31 when no sampling method is used. The maximum test accuracy achieved was 61%. AUC and sensitivity were also very low. Over-sampling reduced the gap between specificity and sensitivity values, however it did not help improve accuracy. The performance of RF and SVM was similar in both sampling approaches.

In case of under-sampling, AUC, sensitivity, and specificity, were all between 0.5 and 0.55. None of the algorithms showed a smooth accuracy curve; the accuracy

Table 3.10: Top 20 features for MCI-NC vs MCI-C

T-test	Relief-F	GiniIndex
log10(CORTISOL)	log10(FSH)	log10(CRP)
log10(Apo_C3)	log10(THBP)	log10(FABP)
log10(ApoE)	log10(Apo_C3)	log10(Apo_C3)
log10(CRP)	log10(ATEN)	log10(CgA)
log10(FABP)	log10(CORTISOL)	log10(IIGFBP)
log10(HGF)	log10(CRP)	log10(ANG2)
log10(NrCAM)	log10(PPP)	log10(ApoE)
log10(SAP)	log10(CD40_L)	log10(FSH)
log10(FSH)	log10(CEA)	log10(CORTISOL)
log10(IIGFBP)	log10(TEST)	log10(LPA)
log10(PAP)	log10(CA_19_9)	log10(MDC)
log10(PARC)	log10(CgA)	log10(IL18)
log10(HBEGF)	log10(EDNA)	log10(Insulin)
log10(CgA)	log10(Insulin)	log10(MIP1alpha)
log10(MMP2)	log10(ApoD)	log10(PAP)
log10(ANG2)	log10(ApoE)	log10(SAP)
log10(ATEN)	log10(AXL)	log10(ATEN)
log10(AXL)	log10(EGF)	log10(I309)
log10(CA_19_9)	log10(IIGFBP)	log10(MMP9)
log10(I309)	log10(LPA)	log10(TEST)
InfoGain	**KruskalWallis**	**SSLogisticR**
log10(A2Micro)	log10(ACE)	log10(CORTISOL)
log10(ACE)	log10(Apo_C3)	log10(FABP)
log10(ADP)	log10(ATEN)	log10(CRP)
log10(AFP)	log10(CORTISOL)	log10(ApoE)
log10(ANG2)	log10(FSH)	log10(Apo_C3)
log10(Apo_A1)	log10(IIGFBP)	log10(SAP)
log10(Apo_C3)	log10(ANG2)	log10(PAP)
log10(ApoB)	log10(Apo_A1)	log10(NrCAM)
log10(ApoD)	log10(PPP)	log10(CgA)
log10(ApoE)	log10(Insulin)	log10(IIGFBP)
log10(ApoH)	log10(PAI1)	log10(HBEGF)
log10(ATEN)	log10(SHBG)	log10(HGF)
log10(AXL)	log10(GH)	log10(LPA)
log10(B2M)	log10(LH)	log10(I309)
log10(BDNF)	log10(AFP)	log10(FSH)
log10(BLC)	log10(MMIF)	log10(MMP2)
log10(BMP_6)	log10(Myoglobin)	log10(AXL)
log10(BTC)	log10(IGA)	log10(IGA)
log10(C3)	log10(ApoB)	log10(FAC7)
log10(CA_19_9)	log10(CD40)	log10(CA_19_9)

dropped after 25 features in all cases. Reilef-F and GiniIndex gave 58% and 57% accuracy respectively for the topmost feature. However, as the number of features were increased, accuracy fluctuated and dropped. Performance of T-Test, FisherKernel, and SSLogisticR was comparable. Kruskal-Wallis gave the highest accuracy of 61.3%, but the accuracy dropped with increasing features. InfoGain, ChiSquare, and mRMR performed better than the aforementioned algorithms. The accuracy obtained was 59%

with increased stability compared to other techniques. InfoGain was the most promising amongst the three. Performance measures of various feature selection algorithms using RF and SVM classifiers are shown in Figures 3.17 and 3.18 respectively. Table 3.10 shows the top performing algorithms with top features.

3.6 ADNI Proteomics: NL vs MCI-NC

The purpose of this experiment was to identify potent biomarkers for predicting NL (negative class) from MCI-NC (positive class) subjects. The dataset setup is similar to Section 3.3. Table 3.11 gives the break-up of training and testing data used in various sampling methods:

Table 3.11: NL vs MCI-NC : Dataset Details

		No	**Sampling**	**Under**	**Sampling**	**Over**	**Sampling**
Target	**Sample #**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
MCI-NC(positive)	218	196	22	48	22	196	22
NL(negative)	54	48	6	48	6	192	6
Total	**165**	**147**	**18**	**96**	**18**	**195**	**18**

3.6.1 Results

Table 3.12: Top 20 Features for NL vs MCI-NC

Relief-F	**GiniIndex**	**InfoGain**	**mRMR**	**SSLogisticR**
log10(HBEGF)	log10(HBEGF)	log10(ACE)	log10(ApoH)	log10(MIP1alpha)
log10(MIP1alpha)	log10(MIP1alpha)	log10(ADP)	log10(ATEN)	log10(HBEGF)
log10(ApoD)	log10(BTC)	log10(AFP)	log10(AXL)	log10(ApoD)
log10(PAPPA)	log10(SGOT)	log10(ANG2)	log10(B2M)	log10(ATEN)
log10(SGOT)	log10(ANG2)	log10(Apo_A1)	log10(BDNF)	log10(PAPPA)
log10(ATEN)	log10(ApoD)	log10(Apo_C3)	log10(BLC)	log10(ACE)
log10(ANG2)	log10(HGF)	log10(ApoB)	log10(BMP_6)	log10(PARC)
log10(FASL)	log10(ATEN)	log10(ApoD)	log10(BTC)	log10(ANG2)
log10(HGF)	log10(EGFR)	log10(ApoE)	log10(C3)	log10(IL16)
log10(IGM)	log10(IGM)	log10(ApoH)	log10(CA_19_9)	log10(IGM)
log10(IL16)	log10(ACE)	log10(ATEN)	log10(CD40)	log10(A2Micro)
log10(A2Micro)	log10(FASL)	log10(AXL)	log10(HBEGF)	log10(TSH)
log10(ACE)	log10(Leptin)	log10(HBEGF)	log10(MIP1alpha)	log10(FASL)
log10(Apo_C3)	log10(PAPPA)	log10(MIP1alpha)	log10(ApoD)	log10(CgA)
log10(BTC)	log10(PPP)	log10(BTC)	log10(ApoB)	log10(RESISTIN)
log10(MMIF)	log10(A2Micro)	log10(SGOT)	log10(CD40_L)	log10(CORTISOL)
log10(RESISTIN)	log10(Apo_C3)	log10(B2M)	log10(SGOT)	log10(Leptin)
log10(TEST)	log10(CRP)	log10(EGFR)	log10(Apo_A1)	log10(SGOT)
log10(CRP)	log10(FAC7)	log10(HGF)	log10(Apo_C3)	log10(MMIF)

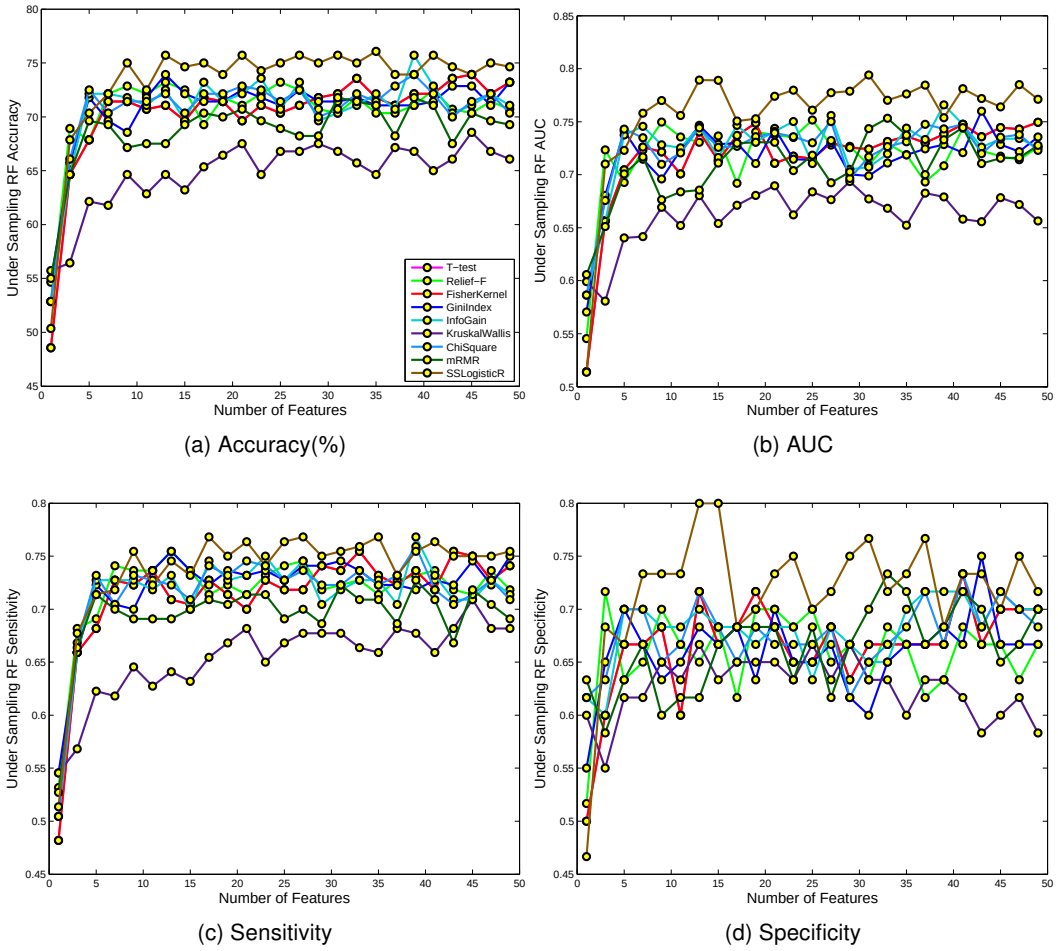
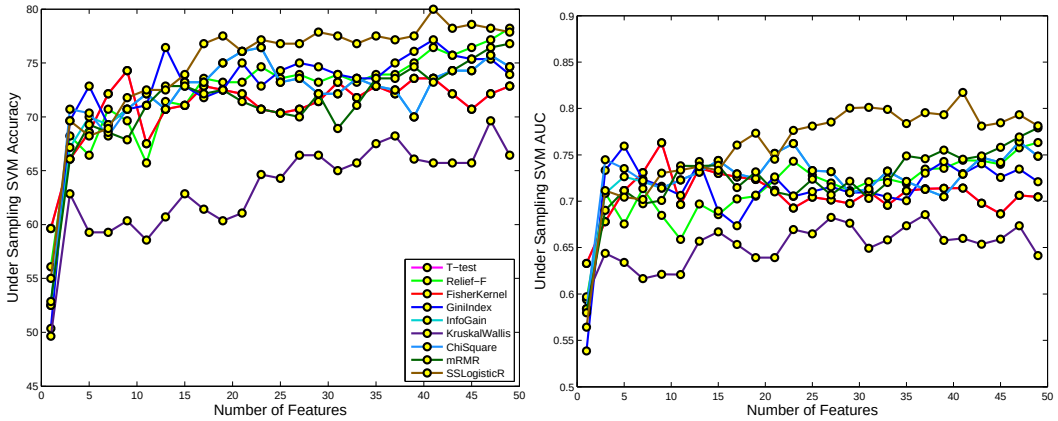


Figure 3.19: NL vs MCI-NC Classification using Under-Sampling features & RF Classifier.

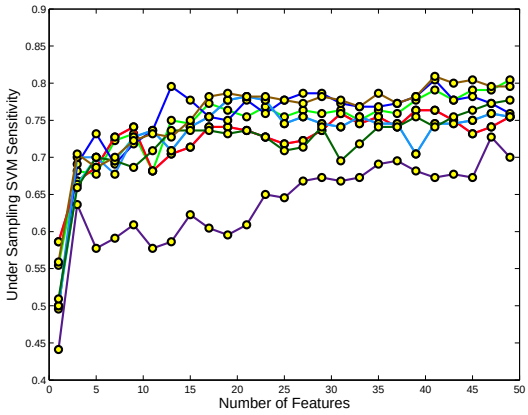
Without using any sampling, RF gave unsatisfactory results with an average sensitivity of 1.0 and a specificity of 0.1. SVM performed better than RF with an average sensitivity of 0.93 and a specificity of 0.45. SSLogisticR, using SVM, showed the best and stable results with 87% accuracy using the top 25 features. Over-sampling approach did not show any improvement; working well for couple of topmost features, but as the number of features increased, sensitivity started increasing, decreasing AUC and specificity. The performance of SVM was better than RF.

Under-sampling approach helped improve the performance measure giving AUC, sensitivity, and specificity between 0.7 and 0.8. The best performing algorithm was SS-

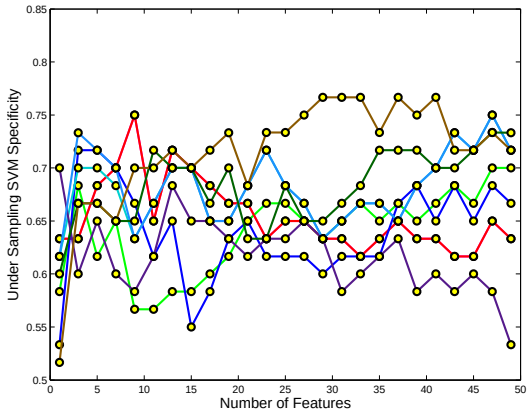


(a) Accuracy(%)

(b) AUC



(c) Sensitivity



(d) Specificity

Figure 3.20: NL vs MCI-NC Classification using Under-Sampling features & SVM Classifier.

LogisticR followed by Relief-F, GiniIndex, mRMR, and InfoGain. ChiSquare was comparable to InfoGain; Kruskal-Wallis did not perform well; T-Test and FisherKernel were average in performance. In RF, the maximum accuracy observed was 77%; whereas SVM reported the highest accuracy of 80.35%. Table 3.12 shows the list of top features obtained using the top algorithms. Performance measures of different feature selection algorithms using RF and SVM classifiers are illustrated in Figures 3.19 and 3.20 respectively.

CHAPTER 4

CONCLUSIONS AND FUTURE WORK

4.1 Conclusions

Alzheimer′s Disease (AD) is a neuro-degenerative disease affecting the elderly. Average life expectancy after diagnosis of the disease is seven years. Most of the treatments aim to provide relief from the symptoms of the disease. It is imperative to diagnose the disease so that future treatments could then target the disease in its earliest stages, before irreversible brain damage or mental decline has occurred. Several genetic, imaging and biochemical markers are being explored to monitor progression of AD and explore treatment and detection options. Identification of blood based biomarkers would offer a quick and effective way to diagnose and monitor the disease. This novel study analyzes Alzheimer′s Disease Neuroimaging Initiative (ADNI) data using numerous integrative approaches to find the most potent biomarkers for early diagnosis of AD. This work differs from previous efforts in its use of a multitude of feature selection (FS) Algorithms for prediction and prognosis of AD.

The FS Algorithms studied in this thesis are unpaired Student′s t-test (T-Test), Relief-F with 10 nearest neighbors, Information Gain, GiniIndex, Chi-Square, FisherKernel, Kruskal-Wallis, mRMR with Mutual Information as the distance measure between the class and the feature, and Logistic regression with ℓ_1-norm regularization and Stability Selection (SSLogisticR). Two metrics have been used to gauge the performance of feature selection algorithms: the relevance of biomarkers with respect to the prior art and the predictive power of the algorithm using Random Forest (RF) and Support Vector Machines (SVM) classifiers. Samples with missing values and outliers were discarded from the study. Various sampling approaches were analyzed to handle unbalanced data.

Sparse Logistic Regression algorithm proved to be most effective. The performance of the algorithm increased with the increasing number of features and it was observed to handle unbalanced data elegantly compared to other algorithms. Relief-F

performed well in most cases giving good results with a small subset of features. T-Test and FisherKernel were similar in performance and were stable compared to Reilef-F. However, both of these techniques required a larger feature subset to achieve the same level of performance as Reilef-F. GiniIndex was comparable to Relief-F. It performed well with smaller subset of features than Relief-F. However these results did not hold for all prediction tasks. InfoGain and ChiSquare performed equally well. Both of these methods, designed for categorical or discrete values, did not generalize well for continuous values. They showed average performance and were not stable with increasing number of features. Kruskal-Wallis did not show satisfactory performance in most prediction tasks except in MCI-NC vs MCI-C prediction, where it was one of the top performing algorithms. Since the algorithm is based on ranking of an observation, which is a discrete value, it did not perform well on continuous value data. mRMR was a computationally expensive algorithm and did not give satisfactory results.

While dealing with unbalanced data, the results indicate that no-sampling approach did not work well due to high bias towards the majority class. Over-sampling approach also did not give satisfactory results. Under-sampling strategy worked well in all prediction tasks. The performances of RF and SVM were comparable.

The key finding of this research is that an integrative approach which uses RBM, MRI, and META together is more effective than using these factors individually. The biomarkers identified by this study have been either linked to AD directly or known to play an important role in other diseases which share the same symptoms as AD.

4.2 Future Work

The main focus of this work was to identify potential protein signatures. This work can be extended to ascertain the robustness of the derived signatures on longitudinal data. In future, multivariate approaches utilizing interactions between the samples and the features can be evaluated for biosignature discovery. The identified proteomics bio-markers can be ported to other datasets which use the same set of features. An

integrative approach combining RBM, MRI, and META biomarkers showed promising results in this work. This approach can be further extended to integrate other biomarkers, such CSF and demographic.

The use of tuning parameters to improve the efficacy of feature selection algorithms can be explored. For example, Relief-F can be made more stable by trying different values of k nearest neighbors on the validation set for each prediction task. RF and SVM parameters can be tuned by cross validation to achieve higher performance. Over and under sampling approaches use random duplication and elimination of the data points respectively. In future, the effect of random addition-deletion can be mitigated by generating more datasets for each task.

BIBLIOGRAPHY

[1] Alzheimer's disease genetics fact sheet. http://www.nia.nih.gov/alzheimers/publication/alzheimers-disease-genetics-fact-sheet.

[2] Australian imaging, biomarker & lifestyle. http://www.aibl.csiro.au/.

[3] Wikipedia web resource. http://www.wikipedia.com.

[4] C. Alzheimer, S. Werner, et al. Fibroblast growth factors and neuroprotection. *Adv Exp Med Biol*, 513:335–351, 2002.

[5] G. Arsenault-Lapierre, V. Whitehead, S. Lupien, and H. Chertkow. Effects of anosognosia on perceived stress and cortisol levels in alzheimerŠs disease. *International Journal of Alzheimer's Disease*, 2012, 2012.

[6] C.M. Bishop and SpringerLink (Service en ligne). *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[7] P. Bongioanni, B. Boccardi, M. Borgna, and B. Rossi. T-lymphocyte interleukin 6 receptor binding in patients with dementia of alzheimer type. *Archives of neurology*, 55(10):1305, 1998.

[8] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] G. Brown, A. Pocock, M.J. Zhao, and M. Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(1):26–66, 2012.

[10] P. Caramelli, R. Nitrini, R. Maranhao, ACG Lourenço, MC Damasceno, C. Vinagre, and B. Caramelli. Increased apolipoprotein b serum concentration in alzheimer's disease. *Acta neurologica scandinavica*, 100(1):61–63, 1999.

[11] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at { http://www.csie.ntu.edu.tw/ cjlin/libsvm }.

[12] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Arxiv preprint arXiv:1106.1813*, 2011.

[13] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[14] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. In *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE*, pages 523–528. IEEE, 2003.

[15] K.A. Ellis, A.I. Bush, D. Darby, D. De Fazio, J. Foster, P. Hudson, N.T. Lautenschlager, N. Lenzo, R.N. Martins, P. Maruff, et al. The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. *International Psychogeriatrics*, 21(04):672–687, 2009.

[16] W.J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, pages 397–416, 1998.

[17] G. Fung and J. Stoeckel. Svm feature selection for classification of spect images of alzheimer's disease using spatial information. *Knowledge and Information Systems*, 11(2):243–258, 2007.

[18] M.V. Gómez-Gaviro, C.E. Scott, A.K. Sesay, A. Matheu, S. Booth, C. Galichet, and R. Lovell-Badge. Betacellulin promotes cell proliferation in the neural stem cell niche and stimulates neurogenesis. *Proceedings of the National Academy of Sciences*, 109(4):1317–1322, 2012.

[19] M.U. Guide. The mathworks. *Inc., Natick, MA*, 5, 1998.

[20] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[21] S. Huang, J. Li, L. Sun, J. Ye, A. Fleisher, T. Wu, K. Chen, E. Reiman, et al. Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–949, 2010.

[22] T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1-2):95–114, 2000.

[23] T.S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999.

[24] A. Jaiantilal. Randomforest-matlab, 2010. http://code.google.com/p/randomforest-matlab/.

[25] D. Johnstone, E.A. Milward, R. Berretta, and P. Moscato. Multivariate protein signatures of pre-clinical alzheimer's disease in the alzheimer's disease neuroimaging initiative (adni) plasma proteome dataset. *PloS one*, 7(4):e34341, 2012.

[26] K. Kira and L.A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the National Conference on Artificial Intelligence*, pages 129–129. John Wiley & Sons Ltd, 1992.

[27] I. Kononenko. Estimating attributes: analysis and extensions of relief. In *Machine Learning: ECML-94*, pages 171–182. Springer, 1994.

[28] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36, 2006.

[29] SB Kotsiantis, D. Kanellopoulos, and PE Pintelas. Data preprocessing for supervised leaning. *International Journal of Computer Science*, 1(2):111–117, 2006.

[30] F. Licastro, I. Carbone, M. Ianni, and E. Porcellini. Gene signature in alzheimer's disease and environmental factors: The virus chronicle. *Journal of Alzheimer's Disease*, 27(4):809–817, 2011.

[31] J. Liu, J. Chen, and J. Ye. Large-scale sparse logistic regression. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–556. ACM, 2009.

[32] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[33] S.J. Lupien, M. De Leon, S. De Santi, A. Convit, C. Tarshish, N.P.V. Nair, M. Thakur, B.S. McEwen, R.L. Hauger, and M.J. Meaney. Cortisol levels during human aging predict hippocampal atrophy and memory deficits. *Nature neuroscience*, 1(1):69–73, 1998.

[34] J.H. McDonald and University of Delaware. *Handbook of biological statistics*. Sparky House Publishing, 2009.

[35] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[36] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C.R. Jack, W. Jagust, J.Q. Trojanowski, A.W. Toga, and L. Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni).

Alzheimer's and Dementia: The Journal of the Alzheimer's Association, 1(1):55–66, 2005.

[37] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[38] M.G. Ravetti and P. Moscato. Identification of a 5-protein biomarker molecular signature for predicting alzheimer's disease. *PloS one*, 3(9):e3111, 2008.

[39] S. Ray, M. Britschgi, C. Herbert, Y. Takeda-Uchimura, A. Boxer, K. Blennow, L.F. Friedman, D.R. Galasko, M. Jutel, A. Karydas, et al. Classification and prediction of clinical alzheimer's diagnosis based on plasma signaling proteins. *Nature medicine*, 13(11):1359–1362, 2007.

[40] M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1):23–69, 2003.

[41] E. Savaskan. The role of the brain renin-angiotensin system in neurodegenerative disorders. *Current Alzheimer Research*, 2(1):29–35, 2005.

[42] J.D. Spurrier. On the null distribution of the kruskal–wallis statistic. *Nonparametric Statistics*, 15(6):685–691, 2003.

[43] D. Tripathy, L. Thirumangalakudi, and P. Grammas. Expression of macrophage inflammatory protein 1-α is elevated in alzheimer's vessels and is regulated by oxidative stress. *Journal of Alzheimer's Disease*, 11(4):447–455, 2007.

[44] Y. Tsuboi, K. Kakimoto, M. Nakajima, H. Akatsu, T. Yamamoto, K. Ogawa, T. Ohnishi, Y. Daikuhara, and T. Yamada. Increased hepatocyte growth factor level in cerebrospinal fluid in alzheimer's disease. *Acta neurologica scandinavica*, 107(2):81–86, 2003.

[45] E. Tuv, A. Borisov, G. Runger, and K. Torkkola. Feature selection with ensembles, artificial variables, and redundancy elimination. *The Journal of Machine Learning Research*, 10:1341–1366, 2009.

[46] K. Une, YA Takei, N. Tomita, T. Asamura, T. Ohrui, K. Furukawa, and H. Arai. Adiponectin in plasma and cerebrospinal fluid in mci and alzheimerŠs disease. *European Journal of Neurology*, 18(7):1006–1009, 2011.

[47] R.A. Whitmer. The epidemiology of adiposity and dementia. *Current Alzheimer Research*, 4(2):117–122, 2007.

[48] J.A. Wood, P.L. Wood, R. Ryan, N.R. Graff-Radford, C. Pilapil, Y. Robitaille, and R. Quirion. Cytokine indices in alzheimer's temporal cortex: no changes in mature il-1 [beta] or il-1ra but increases in the associated acute phase proteins il-6,[alpha] 2-macroglobulin and c-reactive protein. *Brain research*, 629(2):245–252, 1993.

[49] L. Yu and H. Liu. Redundancy based feature selection for microarray data. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 737–742. ACM, 2004.

[50] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu. Advancing feature selection research. *ASU Feature Selection Repository, Arizona State University*, 2010.

APPENDIX A

COMPLETE LIST OF ADNI FEATURES

Table A.1: List of Blood Plasma Protein (RBM) Features

Full Protein Name	Short Name	Full Protein Name	Short Name
Alpha-2-Macroglobulin	A2Micro	factor VII	FAC7
Alpha-1-Antichymotrypsin	AACT	fASLG Receptor	FAS
Alpha-1-Antitrypsin	AAT	fas Ligand	FASL
Angiotensin-Converting enzyme	ACE	fetuin-A	FETA
Adiponectin	ADP	fibroblast Growth factor 4	FGF4
Alpha-fetoprotein	AFP	fibrinogen	FIBGN
Agouti-Related Protein	AGRP	ferritin	FRTN
Angiopoietin-2	ANG2	follicle-Stimulating Hormone	FSH
Apolipoprotein A-I	Apo_A1	Growth Hormone	GH
Apolipoprotein A-II	Apo_A2	Growth-Regulated alpha Protein	GRO_alpha
Apolipoprotein A-IV	Apo_A4	Glutathione S-Transferase alpha	GST_alpha
Apolipoprotein C-I	Apo_C1	Haptoglobin	HAPG
Apolipoprotein C-III	Apo_C3	Heparin-Binding eGf-Like Growth factor	HBEGF
Apolipoprotein B	ApoB	Chemokine CC-4	HCC4
Apolipoprotein D	ApoD	Hepatocyte Growth factor	HGF
Apolipoprotein e	ApoE	T Lymphocyte-Secreted Protein I-309	I309
Apolipoprotein H	ApoH	Intercellular Adhesion Molecule 1	ICAM1
Angiotensinogen	ATEN	Interferon gamma Induced Protein 10	IFGIP10
AXL Receptor Tyrosine kinase	AXL	Immunoglobulin A	IGA
Beta-2-Microglobulin	B2M	Immunoglobulin e	IGE
Brain-Derived neurotrophic factor	BDNF	Immunoglobulin M	IGM
B Lymphocyte Chemoattractant	BLC	Insulin-like Growth factor-Binding Protein 2	IIGFBP
Bone Morphogenetic Protein 6	BMP_6	Interleukin-13	IL13
N-terminal prohormone of brain natriuretic peptide (nT-proBnP)	BNP	Interleukin-16	IL16
Betacellulin	BTC	Interleukin-18	IL18
Complement C3	C3	Interleukin-3	Il3
Cancer Antigen 19-9	CA_19_9	Interleukin-6 Receptor	IL6r
Calcitonin	CALC	Interleukin-8	IL8
CD 40 Antigen	CD40	Insulin	Insulin
CD40 Ligand	CD40_L	kidney Injury Molecule-1	KIM1
CD5 Antigen-like	CD5L	Leptin	Leptin
Carcinoembryonic Antigen	CEA	Luteinizing Hormone	LH
Complement factor H	CFH	Apolipoprotein(a)	LPA
Chromogranin-A	CgA	Monocyte Chemotactic Protein 1	MCP1
Creatine kinase-MB	CK_MB	Monocyte Chemotactic Protein 2	MCP2
Clusterin	CLU	Monocyte Chemotactic Protein 3	MCP3
Ciliary neurotrophic factor	CNTF	Monocyte Chemotactic Protein 4	MCP4
Cortisol	CORTISOL	Macrophage-Derived Chemokine	MDC
C-Peptide	CPEP	Macrophage Colony-Stimulating factor 1	MGSF1
C-Reactive Protein	CRP	Monokine Induced by Gamma Interferon	MIGI
Cystatin-C	CSTC	Macrophage Inflammatory Protein-1 alpha	MIP1alpha
epithelial-Derived neutrophil-Activating Protein 78	EDNA	Macrophage Inflammatory Protein-1 beta	MIP1beta

Continued on next page

epidermal Growth factor	EGF	Macrophage Inflammatory Protein-3 alpha	MIP3alpha
epidermal Growth factor Receptor	EGFR	Macrophage Migration Inhibitory factor	MMIF
e-Selectin	ESEL	Matrix Metalloproteinase-1	MMP1
eotaxin-1	ETAX1	Matrix Metalloproteinase-10	MMP10
eotaxin-3	ETAX3	Matrix Metalloproteinase-2	MMP2
fatty Acid-Binding Protein, Heart	FABP	Matrix Metalloproteinase-7	MMP7
Matrix Metalloproteinase-9	MMP9	Serum Glutamic Oxaloacetic Transaminase	SGOT
Matrix Metalloproteinase-9, Total	MMP9T	Sex Hormone-Binding Globulin	SHBG
Myeloid Progenitor Inhibitory factor 1	MPIF1	Superoxide Dismutase 1, Soluble	SOD1
Myeloperoxidase	MPO	Sortilin	SORTILIN
Myoglobin	Myoglobin	Thyroxine-Binding Globulin	TBG
neutrophil Gelatinase-Associated Lipocalin	NGAL	Thrombospondin-1	TBSP1
neuronal Cell Adhesion Molecule	NrCAM	Thymus-Expressed Chemokine	TECK
Osteopontin	OTP	Testosterone, Total	TEST
Plasminogen Activator Inhibitor 1	PAI1	Trefoil factor 3	TFF3
Prostatic Acid Phosphatase	PAP	Thrombopoietin	THBP
Pregnancy-Associated Plasma Protein A	PAPPA	Tamm-Horsfall urinary Glycoprotein	THP
Pulmonary and Activation-Regulated Chemokine	PARC	Tissue Inhibitor of Metalloproteinases 1	TIMMP1
Platelet-Derived Growth factor BB	PDGFBB	Thrombomodulin	TM
Proinsulin, Intact	PII	Tenascin-C	TNC
Proinsulin, Total	PIT	Tumor necrosis factor alpha	TNFalpha
Placenta Growth factor	PLGF	Tnf-Related Apoptosis-Inducing Ligand Receptor 3	TNFRAILR
Pancreatic Polypeptide	PPP	Tumor necrosis factor Receptor-Like 2	TNFRL2
Prolactin	PRL	Serotransferrin	TRANSFERRIN
Peptide YY	PYY	Thyroid-Stimulating Hormone	TSH
Receptor for Advanced Glycosylation end Products	RAGE	Transthyretin	TTR
T-Cell-Specific Protein RAnTeS	RANTES	Vascular Cell Adhesion Molecule-1	VCAM
Resistin	RESISTIN	Vascular endothelial Growth factor	VEGF
Serum Amyloid P-Component	SAP	Vitronectin	VITRN
Stem Cell factor	SCF	Vitamin k-Dependent Protein S	VKDPS
		von Willebrand factor	vWF

Table A.2: List of Reduced Proteomics Features

log10(A2Micro)	log10(CA_19_9)	log10(GH)	log10(LH)	log10(SAP)
log10(ACE)	log10(CD40)	log10(GRO_alpha)	log10(LPA)	log10(SCF)
log10(ADP)	log10(CD40_L)	log10(HAPG)	log10(MCP1)	log10(SGOT)
log10(AFP)	log10(CEA)	log10(HBEGF)	log10(MDC)	log10(SHBG)
log10(ANG2)	log10(CFH)	log10(HCC4)	log10(MIP1alpha)	log10(SOD1)
log10(Apo_A1)	log10(CgA)	log10(HGF)	log10(MIP1beta)	log10(SORTILIN)
log10(Apo_C3)	log10(CK_MB)	log10(I309)	log10(MMIF)	log10(TBG)
log10(ApoB)	log10(CORTISOL)	log10(ICAM1)	log10(MMP2)	log10(TBSP1)
log10(ApoD)	log10(CRP)	log10(IGA)	log10(MMP9)	log10(TECK)
log10(ApoE)	log10(EDNA)	log10(IGE)	log10(MPO)	log10(TEST)
log10(ApoH)	log10(EGF)	log10(IGM)	log10(Myoglobin)	log10(THBP)
log10(ATEN)	log10(EGFR)	log10(IIGFBP)	log10(NrCAM)	log10(TIMMP1)
log10(AXL)	log10(ETAX1)	log10(IL13)	log10(PAI1)	log10(TNC)
log10(B2M)	log10(FABP)	log10(IL16)	log10(PAP)	log10(TNFRAILR)
log10(BDNF)	log10(FAC7)	log10(IL18)	log10(PAPPA)	log10(TNFRL2)
log10(BLC)	log10(FASL)	log10(Il3)	log10(PARC)	log10(TSH)
log10(BMP_6)	log10(FIBGN)	log10(IL8)	log10(PPP)	log10(VCAM)
log10(BTC)	log10(FRTN)	log10(Insulin)	log10(PRL)	log10(VEGF)
log10(C3)	log10(FSH)	log10(Leptin)	log10(RESISTIN)	log10(vWF)

Table A.3: List of Psychometric Assessment Scores (META Features)

NPI	Test RCT4; ALT (SGPT)
ANARTERR	Test RCT5; AST (SGOT)
BNTTOTAL	Test RCT6; Urea Nitrogen
CATANIMSC	Test RCT8; Serum Uric Acid
CATVEGESC	Test RCT9; Phosphorus
CLOCKSCOR	ADAS_sub1
DIGITSCOR	ADAS_sub2
DSPANBAC	ADAS_sub3
TRAASCOR	ADAS_sub4
TRABSCOR	ADAS_sub5
MMSE	ADAS_sub6
LDELTOTAL	ADAS_sub7
LIMMTOTAL	ADAS_sub8
Test RCT1; Total Bilirubin	ADAS_sub9
Test RCT11; Serum Glucose	ADAS_sub10
Test RCT12; Total Protein	ADAS_sub11
Test RCT13; Albumin	ADAS_sub12
Test RCT14; Creatine Kinase	ADAS_sub13
Test RCT1407; Alkaline Phosphatase	Hachinski
Test RCT1408; LDH	GDS
Test RCT183; Calcium (EDTA)	FAQ
Test RCT19; Triglycerides (GPO)	Age
Test RCT20; Cholesterol (High Performance)	Gender
Test RCT29; Direct Bilirubin	Educat
Test RCT3; GGT	CDR
Test RCT392; Creatinine (Rate Blanked)	APOE

Table A.4: List of Magnetic Resonance Imaging (MRI) Features

Volume (WM Parcellation) of RightPallidum	Volume (WM Parcellation) of CorpusCallosumCentral
Volume (Cortical Parcellation) of RightParacentral	Volume (Cortical Parcellation) of LeftMiddleTemporal
Surface Area of RightParacentral	Surface Area of LeftMiddleTemporal
Cortical Thickness Avg of RightParacentral	Cortical Thickness Avg of LeftMiddleTemporal
Cortical Thickness STDV of RightParacentral	Cortical Thickness STDV of LeftMiddleTemporal
Volume (Cortical Parcellation) of RightParahippocampal	Volume (WM Parcellation) of LeftPallidum
Surface Area of RightParahippocampal	Volume (Cortical Parcellation) of LeftParacentral
Cortical Thickness Avg of RightParahippocampal	Surface Area of LeftParacentral
Cortical Thickness STDV of RightParahippocampal	Cortical Thickness Avg of LeftParacentral
Volume (Cortical Parcellation) of RightParsOpercularis	Cortical Thickness STDV of LeftParacentral
Surface Area of RightParsOpercularis	Volume (Cortical Parcellation) of LeftParahippocampal
Cortical Thickness Avg of RightParsOpercularis	Surface Area of LeftParahippocampal
Cortical Thickness STDV of RightParsOpercularis	Cortical Thickness Avg of LeftParahippocampal
Volume (Cortical Parcellation) of RightParsOrbitalis	Cortical Thickness STDV of LeftParahippocampal
Surface Area of RightParsOrbitalis	Volume (Cortical Parcellation) of LeftParsOpercularis
Cortical Thickness Avg of RightParsOrbitalis	Surface Area of LeftParsOpercularis
Cortical Thickness STDV of RightParsOrbitalis	Cortical Thickness Avg of LeftParsOpercularis
Volume (Cortical Parcellation) of RightParsTriangularis	Cortical Thickness STDV of LeftParsOpercularis
Surface Area of RightParsTriangularis	Volume (Cortical Parcellation) of LeftParsOrbitalis
Cortical Thickness Avg of RightParsTriangularis	Surface Area of LeftParsOrbitalis
Cortical Thickness STDV of RightParsTriangularis	Cortical Thickness Avg of LeftParsOrbitalis
Volume (Cortical Parcellation) of RightPericalcarine	Cortical Thickness STDV of LeftParsOrbitalis
Surface Area of RightPericalcarine	Volume (Cortical Parcellation) of LeftParsTriangularis
Cortical Thickness Avg of RightPericalcarine	Surface Area of LeftParsTriangularis
Cortical Thickness STDV of RightPericalcarine	Cortical Thickness Avg of LeftParsTriangularis
Volume (Cortical Parcellation) of RightPostcentral	Cortical Thickness STDV of LeftParsTriangularis
Surface Area of RightPostcentral	Volume (Cortical Parcellation) of LeftPericalcarine
Cortical Thickness Avg of RightPostcentral	Surface Area of LeftPericalcarine
Cortical Thickness STDV of RightPostcentral	Cortical Thickness Avg of LeftPericalcarine
Volume (Cortical Parcellation) of RightPosteriorCingulate	Cortical Thickness STDV of LeftPericalcarine
Surface Area of RightPosteriorCingulate	Volume (Cortical Parcellation) of LeftPostcentral
Cortical Thickness Avg of RightPosteriorCingulate	Surface Area of LeftPostcentral
Cortical Thickness STDV of RightPosteriorCingulate	Cortical Thickness Avg of LeftPostcentral
Volume (Cortical Parcellation) of Icv	Cortical Thickness STDV of LeftPostcentral
Volume (Cortical Parcellation) of RightPrecentral	Volume (WM Parcellation) of CorpusCallosumMidAnterior
Surface Area of RightPrecentral	Volume (Cortical Parcellation) of LeftPosteriorCingulate
Cortical Thickness Avg of RightPrecentral	Surface Area of LeftPosteriorCingulate
Cortical Thickness STDV of RightPrecentral	Cortical Thickness Avg of LeftPosteriorCingulate

Continued on next page

Volume (Cortical Parcellation) of RightPrecuneus	Cortical Thickness STDV of LeftPosteriorCingulate
Surface Area of RightPrecuneus	Volume (Cortical Parcellation) of LeftPrecentral
Cortical Thickness Avg of RightPrecuneus	Surface Area of LeftPrecentral
Cortical Thickness STDV of RightPrecuneus	Cortical Thickness Avg of LeftPrecentral
Volume (WM Parcellation) of RightPutamen	Cortical Thickness STDV of LeftPrecentral
Volume (Cortical Parcellation) of RightRostralAnteriorCingulate	Volume (Cortical Parcellation) of LeftPrecuneus
Surface Area of RightRostralAnteriorCingulate	Surface Area of LeftPrecuneus
Cortical Thickness Avg of RightRostralAnteriorCingulate	Cortical Thickness Avg of LeftPrecuneus
Cortical Thickness STDV of RightRostralAnteriorCingulate	Cortical Thickness STDV of LeftPrecuneus
Volume (Cortical Parcellation) of RightRostralMiddleFrontal	Volume (WM Parcellation) of LeftPutamen
Surface Area of RightRostralMiddleFrontal	Volume (Cortical Parcellation) of LeftRostralAnteriorCingulate
Cortical Thickness Avg of RightRostralMiddleFrontal	Surface Area of LeftRostralAnteriorCingulate
Cortical Thickness STDV of RightRostralMiddleFrontal	Cortical Thickness Avg of LeftRostralAnteriorCingulate
Volume (Cortical Parcellation) of RightSuperiorFrontal	Cortical Thickness STDV of LeftRostralAnteriorCingulate
Surface Area of RightSuperiorFrontal	Volume (Cortical Parcellation) of LeftRostralMiddleFrontal
Cortical Thickness Avg of RightSuperiorFrontal	Surface Area of LeftRostralMiddleFrontal
Cortical Thickness STDV of RightSuperiorFrontal	Cortical Thickness Avg of LeftRostralMiddleFrontal
Volume (Cortical Parcellation) of RightSuperiorParietal	Cortical Thickness STDV of LeftRostralMiddleFrontal
Surface Area of RightSuperiorParietal	Volume (Cortical Parcellation) of LeftSuperiorFrontal
Cortical Thickness Avg of RightSuperiorParietal	Surface Area of LeftSuperiorFrontal
Cortical Thickness STDV of RightSuperiorParietal	Cortical Thickness Avg of LeftSuperiorFrontal
Volume (Cortical Parcellation) of RightSuperiorTemporal	Cortical Thickness STDV of LeftSuperiorFrontal
Surface Area of RightSuperiorTemporal	Volume (Cortical Parcellation) of LeftSuperiorParietal
Cortical Thickness Avg of RightSuperiorTemporal	Surface Area of LeftSuperiorParietal
Cortical Thickness STDV of RightSuperiorTemporal	Cortical Thickness Avg of LeftSuperiorParietal
Volume (Cortical Parcellation) of RightSupramarginal	Cortical Thickness STDV of LeftSuperiorParietal
Surface Area of RightSupramarginal	Volume (Cortical Parcellation) of LeftSuperiorTemporal
Cortical Thickness Avg of RightSupramarginal	Surface Area of LeftSuperiorTemporal
Cortical Thickness STDV of RightSupramarginal	Cortical Thickness Avg of LeftSuperiorTemporal
Volume (Cortical Parcellation) of RightTemporalPole	Cortical Thickness STDV of LeftSuperiorTemporal
Surface Area of RightTemporalPole	Volume (Cortical Parcellation) of LeftSupramarginal
Cortical Thickness Avg of RightTemporalPole	Surface Area of LeftSupramarginal
Cortical Thickness STDV of RightTemporalPole	Cortical Thickness Avg of LeftSupramarginal
Volume (WM Parcellation) of RightThalamus	Cortical Thickness STDV of LeftSupramarginal
Volume (Cortical Parcellation) of RightTransverseTemporal	Volume (Cortical Parcellation) of LeftTemporalPole
Surface Area of RightTransverseTemporal	Surface Area of LeftTemporalPole

Continued on next page

Cortical Thickness Avg of RightTransverseTemporal	Cortical Thickness Avg of LeftTemporalPole
Cortical Thickness STDV of RightTransverseTemporal	Cortical Thickness STDV of LeftTemporalPole
Volume (WM Parcellation) of RightVentralDC	Volume (WM Parcellation) of LeftThalamus
Volume (WM Parcellation) of ThirdVentricle	Volume (Cortical Parcellation) of LeftTransverseTemporal
Volume (Cortical Parcellation) of LeftInsula	Surface Area of LeftTransverseTemporal
Surface Area of LeftInsula	Cortical Thickness Avg of LeftTransverseTemporal
Cortical Thickness Avg of LeftInsula	Cortical Thickness STDV of LeftTransverseTemporal
Cortical Thickness STDV of LeftInsula	Volume (WM Parcellation) of LeftVentralDC
Volume (WM Parcellation) of LeftAmygdala	Volume (WM Parcellation) of OpticChiasm
Volume (Cortical Parcellation) of RightInsula	Volume (WM Parcellation) of RightAmygdala
Surface Area of RightInsula	Volume (Cortical Parcellation) of RightBankssts
Cortical Thickness Avg of RightInsula	Surface Area of RightBankssts
Cortical Thickness STDV of RightInsula	Cortical Thickness Avg of RightBankssts
Volume (Cortical Parcellation) of LeftBankssts	Cortical Thickness STDV of RightBankssts
Surface Area of LeftBankssts	Volume (Cortical Parcellation) of RightCaudalAnteriorCingulate
Cortical Thickness Avg of LeftBankssts	Surface Area of RightCaudalAnteriorCingulate
Cortical Thickness STDV of LeftBankssts	Cortical Thickness Avg of RightCaudalAnteriorCingulate
Volume (Cortical Parcellation) of LeftCaudalAnteriorCingulate	Cortical Thickness STDV of RightCaudalAnteriorCingulate
Surface Area of LeftCaudalAnteriorCingulate	Volume (Cortical Parcellation) of RightCaudalMiddleFrontal
Cortical Thickness Avg of LeftCaudalAnteriorCingulate	Surface Area of RightCaudalMiddleFrontal
Cortical Thickness STDV of LeftCaudalAnteriorCingulate	Cortical Thickness Avg of RightCaudalMiddleFrontal
Volume (Cortical Parcellation) of LeftCaudalMiddleFrontal	Cortical Thickness STDV of RightCaudalMiddleFrontal
Surface Area of LeftCaudalMiddleFrontal	Volume (WM Parcellation) of RightCaudate
Cortical Thickness Avg of LeftCaudalMiddleFrontal	Volume (WM Parcellation) of RightCerebellumCortex
Cortical Thickness STDV of LeftCaudalMiddleFrontal	Volume (WM Parcellation) of RightCerebellumWM
Volume (WM Parcellation) of LeftCaudate	Volume (WM Parcellation) of RightCerebralCortex
Volume (WM Parcellation) of LeftCerebellumCortex	Volume (WM Parcellation) of RightCerebralWM
Volume (WM Parcellation) of LeftCerebellumWM	Volume (WM Parcellation) of Csf
Volume (WM Parcellation) of LeftCerebralCortex	Volume (WM Parcellation) of RightChoroidPlexus
Volume (WM Parcellation) of Brainstem	Volume (Cortical Parcellation) of RightCuneus
Volume (WM Parcellation) of LeftCerebralWM	Surface Area of RightCuneus
Volume (WM Parcellation) of LeftChoroidPlexus	Cortical Thickness Avg of RightCuneus
Volume (Cortical Parcellation) of LeftCuneus	Cortical Thickness STDV of RightCuneus
Surface Area of LeftCuneus	Surface Area of RightEntorhinal
Cortical Thickness Avg of LeftCuneus	Cortical Thickness Avg of RightEntorhinal
Cortical Thickness STDV of LeftCuneus	Cortical Thickness STDV of RightEntorhinal
Volume (Cortical Parcellation) of LeftEntorhinal	Volume (Cortical Parcellation) of RightFrontalPole
Surface Area of LeftEntorhinal	Surface Area of RightFrontalPole
Cortical Thickness Avg of LeftEntorhinal	Cortical Thickness Avg of RightFrontalPole
Cortical Thickness STDV of LeftEntorhinal	Cortical Thickness STDV of RightFrontalPole
Volume (Cortical Parcellation) of LeftFrontalPole	Volume (Cortical Parcellation) of RightFusiform

Continued on next page

Surface Area of LeftFrontalPole	Surface Area of RightFusiform
Cortical Thickness Avg of LeftFrontalPole	Cortical Thickness Avg of RightFusiform
Cortical Thickness STDV of LeftFrontalPole	Cortical Thickness STDV of RightFusiform
Volume (Cortical Parcellation) of LeftFusiform	Surface Area of RightHemisphere
Surface Area of LeftFusiform	Volume (WM Parcellation) of RightHippocampus
Cortical Thickness Avg of LeftFusiform	Volume (Cortical Parcellation) of RightInferiorParietal
Cortical Thickness STDV of LeftFusiform	Surface Area of RightInferiorParietal
Surface Area of LeftHemisphere	Cortical Thickness Avg of RightInferiorParietal
Volume (WM Parcellation) of LeftHippocampus	Cortical Thickness STDV of RightInferiorParietal
Volume (Cortical Parcellation) of LeftInferiorParietal	Volume (Cortical Parcellation) of RightInferiorTemporal
Surface Area of LeftInferiorParietal	Surface Area of RightInferiorTemporal
Cortical Thickness Avg of LeftInferiorParietal	Cortical Thickness Avg of RightInferiorTemporal
Cortical Thickness STDV of LeftInferiorParietal	Cortical Thickness STDV of RightInferiorTemporal
Volume (Cortical Parcellation) of LeftInferiorTemporal	Volume (Cortical Parcellation) of RightIsthmusCingulate
Surface Area of LeftInferiorTemporal	Surface Area of RightIsthmusCingulate
Cortical Thickness Avg of LeftInferiorTemporal	Cortical Thickness Avg of RightIsthmusCingulate
Cortical Thickness STDV of LeftInferiorTemporal	Cortical Thickness STDV of RightIsthmusCingulate
Volume (Cortical Parcellation) of LeftIsthmusCingulate	Volume (Cortical Parcellation) of RightLateralOccipital
Surface Area of LeftIsthmusCingulate	Surface Area of RightLateralOccipital
Cortical Thickness Avg of LeftIsthmusCingulate	Cortical Thickness Avg of RightLateralOccipital
Cortical Thickness STDV of LeftIsthmusCingulate	Cortical Thickness STDV of RightLateralOccipital
Volume (Cortical Parcellation) of LeftLateralOccipital	Volume (Cortical Parcellation) of RightLateralOrbitofrontal
Surface Area of LeftLateralOccipital	Surface Area of RightLateralOrbitofrontal
Cortical Thickness Avg of LeftLateralOccipital	Cortical Thickness Avg of RightLateralOrbitofrontal
Cortical Thickness STDV of LeftLateralOccipital	Cortical Thickness STDV of RightLateralOrbitofrontal
Volume (Cortical Parcellation) of LeftLateralOrbitofrontal	Volume (Cortical Parcellation) of RightLingual
Surface Area of LeftLateralOrbitofrontal	Surface Area of RightLingual
Cortical Thickness Avg of LeftLateralOrbitofrontal	Cortical Thickness Avg of RightLingual
Cortical Thickness STDV of LeftLateralOrbitofrontal	Cortical Thickness STDV of RightLingual
Volume (Cortical Parcellation) of LeftLingual	Volume (Cortical Parcellation) of RightMedialOrbitofrontal
Surface Area of LeftLingual	Surface Area of RightMedialOrbitofrontal
Cortical Thickness Avg of LeftLingual	Cortical Thickness Avg of RightMedialOrbitofrontal
Cortical Thickness STDV of LeftLingual	Cortical Thickness STDV of RightMedialOrbitofrontal
Volume (Cortical Parcellation) of LeftMedialOrbitofrontal	Volume (Cortical Parcellation) of RightMiddleTemporal
Surface Area of LeftMedialOrbitofrontal	Surface Area of RightMiddleTemporal
Cortical Thickness Avg of LeftMedialOrbitofrontal	Cortical Thickness Avg of RightMiddleTemporal
Cortical Thickness STDV of LeftMedialOrbitofrontal	Cortical Thickness STDV of RightMiddleTemporal
	Volume (WM Parcellation) of FourthVentricle