Published in final edited form as:
Neuroimage. 2014 February 15; 87: 220–241. doi:10.1016/j.neuroimage.2013.10.005.

ANALYSIS OF SAMPLING TECHNIQUES FOR IMBALANCED DATA: AN N=648 ADNI STUDY

Rashmi Dubey, MS[1,2], **Jiayu Zhou, BS**[1,2], **Yalin Wang, PhD**[1], **Paul M. Thompson, PhD**[3], and **Jieping Ye, PhD**[1,2] **For the Alzheimer's Disease Neuroimaging Initiative***

[1]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA

[2]Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, Tempe, AZ, USA

[3]Imaging Genetics Center, Laboratory of Neuro Imaging, UCLA School of Medicine, Los Angeles, CA, USA

Abstract

Many neuroimaging applications deal with imbalanced imaging data. For example, in Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, the mild cognitive impairment (MCI) cases eligible for the study are nearly two times the Alzheimer's disease (AD) patients for structural magnetic resonance imaging (MRI) modality and six times the control cases for proteomics modality. Constructing an accurate classifier from imbalanced data is a challenging task. Traditional classifiers that aim to maximize the overall prediction accuracy tend to classify all data into the majority class. In this paper, we study an ensemble system of feature selection and data sampling for the class imbalance problem. We systematically analyze various sampling techniques by examining the efficacy of different rates and types of undersampling, oversampling, and a combination of over and under sampling approaches. We thoroughly examine six widely used feature selection algorithms to identify significant biomarkers and thereby reduce the complexity of the data. The efficacy of the ensemble techniques is evaluated using two different classifiers including Random Forest and Support Vector Machines based on classification accuracy, area under the receiver operating characteristic curve (AUC), sensitivity, and specificity measures. Our extensive experimental results show that for various problem settings in ADNI, **(1).** a balanced training set obtained with K-Medoids technique based undersampling gives the best overall performance among different data sampling techniques and no sampling approach; and **(2).** sparse logistic regression with stability selection achieves competitive performance among various feature selection algorithms. Comprehensive experiments with various settings show that our proposed ensemble model of multiple undersampled datasets yields stable and promising results.

Please address correspondence to: Dr. Jieping Ye, Department of Computer Science and Engineering, Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, 699 S. Mill Ave, Tempe, AZ 85287, jieping.ye@asu.edu.

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Keywords

Alzheimer's disease; classification; imbalanced data; undersampling; oversampling; feature selection

1. INTRODUCTION

Alzheimer's disease (AD) is the most frequent form of dementia in elderly patients; it is a neurodegenerative disease which causes irreversible damage to motor neurons and their connectivity, resulting in cognitive failure and several other behavioral disorders which severely impact day-to-day functioning of the patients (Alzheimer's Association, 2012). As the population is aging, by the year 2050, it is projected that there will be 13.5 million clinical AD individuals accounting for a total care cost of $1.1 trillion (Alzheimer's Association, 2012). It is estimated that by the time the typical patient is diagnosed with AD, the disease has been progressing for nearly a decade. Preclinical AD patients may not show debilitating AD symptoms but the toxic changes in the brain and blood proteins have been developing since inception of the disease (Vlkolinsk et al., 2001; Bartzokis, 2004). Early diagnosis of AD is critical to prevent or delay the progression of the disease. Future treatments could then target the disease in its earliest stages, before irreversible brain damage or mental decline has occurred.

There are many studies which aim to capture the elusive biomarkers of AD for preclinical AD research (Sperling et al., 2011). Several genetic, imaging and biochemical markers are being studied to monitor progression of AD and explore treatment and detection options (Mueller et al., 2005; Jack et al., 2008; Shaw et al., 2009; Frisoni et al., 2010; Reiman and Jagust, 2011). For example, a genetic risk factor, Apolipoprotein E (APOE) gene, has been shown to be associated with the late onset of AD. The APOE gene comes in different forms or alleles; people with an APOE ε -4 allele have a 20% to 90% higher risk of developing Alzheimer's disease than those who do not have an APOE ε -4 (Corder et al., 1993; Mayeux et al., 1998). Magnetic resonance imaging (MRI) and fluorodeoxyglucose positron emission tomography (FDG-PET) scans are powerful neuroimaging modalities which have been shown by various cross-sectional and longitudinal studies to have the highest diagnostic and prognostic power in identifying preclinical and clinical AD patients from control cases (Dickerson et al., 2001; Devanand et al., 2007). MRI is a medical imaging technique utilizing magnetic field to produce very clear 3-dimensional images enabling detailed study of structural and functional changes in the body. MRI has become an essential tool in AD research due to its non-invasive nature, widespread availability, and great potential in predicting disease progression. Since the brain controls most functions of the body, it is hypothesized that any changes in the brain are reflected in the proteins produced. Proteomics, the study of proteins found in blood, is gaining momentum as an AD modality due to its cost effectiveness, ease of availability, and ability to detect probable/positive AD cases in simplistic initial screenings which could be followed up by other advanced clinical modalities (Ray et al., 2007; O'Bryant et al., 2011).

The Alzheimer's Disease Neuroimaging Initiative (ADNI), a multi-pronged, longitudinal study started as a 5 year project, is a collaborative effort by multiple research groups from both the public and private sectors, including the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), 13 pharmaceutical companies, and 2 foundations that provided support through the Foundation for the National Institutes of Health (NIH). It was launched in 2003 as a $60 million, 5-year public-private partnership to help identify the combination of biomarkers with the highest diagnostic and prognostic power. The primary goal of ADNI

has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. This initiative has helped develop optimized methods and uniform standards for acquiring biomarker data which includes MRI, PET, proteomics and genetics data on patients with AD, mild cognitive impairment (MCI) and healthy controls (NC), and creating an accessible data repository for the scientific community (Mueller et al., 2005). The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco.

One of the key challenges in designing good prediction models on ADNI data lies in the class imbalance problem. A dataset is said to be *imbalanced* if there are significantly more data points of one class and fewer occurrences of the other class. For example, the number of control cases in the ADNI dataset is half of the number of AD cases for proteomics measurement, whereas for MRI modality, there are 40% more control cases than AD cases. Data imbalance is also ubiquitous in worldwide ADNI type initiatives from Europe, Japan and Australia, etc. (Weiner et al., 2012). In addition, lots of medical research involves dealing with rare, but important medical conditions/events or subject dropouts in the longitudinal study (Duchesnay et al., 2011; Fitzmaurice et al., 2011; Jiang et al., 2011; Bernal-Rusiel et al., 2012; Johnstone et al., 2012). It is commonly agreed that imbalanced datasets adversely impact the performance of the classifiers as the learned model is biased towards the majority class to minimize the overall error rate (Estabrooks, 2000; Japkowicz, 2000a). For example, in Cuingnet, et al. (2011), due to the imbalance in the number of subjects in NC and MCIc (MCI Converter) groups, they achieved a much lower sensitivity than specificity. Similarly, in our prior work (Yuan et al., 2012), due to the imbalance in the number of subjects in NC, MCI and AD groups, we obtained imbalanced sensitivity and specificity on AD/MCI and MCI/NC classification experiments. Recently, Johnstone et al. (2012) studied pre-clinical AD prediction using proteomics features in the ADNI dataset. They experimented with imbalanced and balanced datasets and observed that the sensitivity and specificity gap significantly reduces when the training set is balanced.

In machine learning field, many approaches have been developed in the past to deal with the imbalanced data (Chan and Stolfo, 1998; Provost, 2000; Japkowicz and Stephen, 2002; Chawla et al., 2003; Ko cz et al., 2003; Maloof, 2003; Chawla et al., 2004; Jo and Japkowicz, 2004; Lee et al., 2004; Visa and Ralescu, 2005; Yang and Wu, 2006; Ertekin et al., 2007; Van Hulse et al., 2007; He and Garcia, 2009; Liu et al., 2009c). They can be broadly classified as internal or algorithmic level and external or data level. The *algorithmic level approaches* involve either designing new classification algorithms or modifying the existing ones to handle the bias introduced due to the class imbalance. Many researchers studied the class imbalance problem in relation to the cost-sensitive learning problem, wherein the penalty of misclassification is different for different class instances, and proposed solutions to the class imbalance problem by increasing the misclassification cost of the minority class and/or by adjusting the estimate at leaf nodes in case of decision trees such as Random Forest (RF) (Knoll et al., 1994; Pazzani et al., 1994; Bradford et al., 1998; Elkan, 2001; Chen et al., 2004). Akbani et al. proposed an algorithm for learning from imbalanced data in case of Support Vector Machines (SVM) by updating the kernel function (Akbani et al., 2004). Recognition based (one-class) learning was identified as a better solution for certain imbalanced datasets instead of two-class learning approaches (Japkowicz, 2001). The *external or data level* solutions include different types of data resampling techniques such as undersampling and oversampling. Random resampling

techniques randomly select data points to be replicated (oversampling with or without replacement) or removed (undersampling). These approaches incur the cost of over-fitting or losing the important information respectively. Directed or focused sampling techniques select specific data points to replicate or remove. Japkowicz proposed to resample minority class instances lying close to the class boundary (Japkowicz, 2000b) whereas Kubat and Matwin (1997) proposed resampling majority class such that borderline and noisy data points are eliminated from the selection. Yen and Lee (2006) proposed cluster-based under-sampling approaches for selecting the representative data as training data to improve the classification accuracy. Liu et al. (2009) developed two ensemble learning systems to overcome the deficiency of information loss introduced in the traditional random undersampling method. Chawla et al. (2002) designed a sophisticated algorithm based on nearest neighbors to generate synthetic data for oversampling (SMOTE) and combined it with undersampling approaches and achieved significant improvements over random sampling techniques. Padmaja et al. (2008) proposed an algorithm, called Majority filter-based minority prediction (MFMP), and achieved better performance than random resampling approaches. Estabrooks et al. (2004) dealt with the rate of resampling required and proposed a combination scheme heavily biased towards under-represented class to mitigate the classifier's bias towards the majority class. Joshi et al. (2001) combined results from several weak classifiers and concluded that boosting algorithms such as RareBoost and AdaBoost effectively handle rare cases. Zheng and Srihari (2003) proposed a novel feature level solution based on selecting and optimally combining positive and negative features. This approach was specifically devised to solve the imbalanced data problem in text categorization.

Apart from the internal and external solutions, evaluation of the classifier for imbalanced datasets has always remained a big challenge (Elkan, 2003). Provost and Fawcett (2001) proposed the ROC convex hull method for estimating classifier performance. Ling and Li (1998) used lift analysis as the performance measure, for marketing analysis problem, which is a customized version of ROC curve. Kubat and Matwin (1997) used the geometric mean to assess the classifier performance. The internal approaches are quite effective; for example, Zadronzy et al. (2003) proposed a cost-sensitive ensemble classifier *Costing* which yielded better results than random sampling methods. However, the greatest disadvantage of internal level solutions is that they are very specific to the classification algorithm. On the other hand, the external or data level solutions are classifier independent, portable, and therefore more adaptable. In this work, we focus on developing and evaluating ensemble models based on data level methods.

While ubiquitous and important, imbalanced data analysis has not received enough attention in the neuroimaging field, at least for the ADNI dataset. This paper aims to fill this gap by studying an ensemble technique to tackle the class imbalance problem in the ADNI dataset. The resampling approaches that we studied include random undersampling and oversampling (Jo and Japkowicz, 2004; Yen and Lee, 2006; Van Hulse et al., 2007; He and Garcia, 2009; Liu et al., 2009c), SMOTE oversampling (Chawla et al., 2002), and K-Medoids based undersampling. We extended our study by varying rates of undersampling and oversampling independently, and a combination of different rates of oversampling and undersampling to generate balanced training sets. In AD research, it is crucial to determine a few significant bio-markers that can help develop therapeutic treatment. In this paper, we examine six state-of-the-art feature selection algorithms including Student's *t*-test, Relief-F, Gini Index, Information Gain, Chi-Square, and Sparse Logistic Regression with stability selection. The classifiers studied are decision tree based Random Forest (RF) classifier and decision boundary based Support Vector Machine (SVM) classifier. The classification evaluation criterion is a combination of test accuracy, AUC, sensitivity, and specificity. As an illustration, we study clinical group (diagnostic) classification problems using the ADNI

baseline MR imaging and proteomics data. The multitude of experiments conducted corroborated the efficacy of the ensemble system which includes an ensemble of multiple completely undersampled datasets (majority class is reduced to match minority class count) using K-Medoids together with feature selection based on sparse logistic regression and stability selection.

2. SUBJECTS AND METHODS

2.1. Subjects

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, including approximately 200 cognitively normal older individuals, 400 people with MCI, and 200 people with early AD. For up-to-date information, see www.adni-info.org.

In our experiments, we used the baseline MRI and proteomics data as the input features because of their wide availability. The MRI image features in this study were based on the imaging data from the ADNI database processed by the UCSF team, who performed cortical reconstruction and volumetric segmentations with the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/). The processed MRI features come from a total of 648 subjects (138AD, 319 MCI and 191 NC), and can be grouped into 5 categories: average cortical thickness, standard deviation in cortical thickness, the volumes of cortical parcellations (based on regions of interest automatically segmented in the cortex), the volumes of specific white matter parcellations, and the total surface area of the cortex. There were 305 MRI features in total. Details of the analysis procedure are available at http://adni.loni.ucla.edu/research/mri-post-processing/. More details on ADNI MRI imaging instrumentation and procedures (Jack et al., 2008) may be found at the ADNI web site (http://adni.loni.ucla.edu). The proteomics data set (112 AD, 396 MCI, and 58 NC) was produced by the Biomarkers Consortium Project "Use of Targeted Multiplex Proteomic Strategies to Identify Plasma-Based Biomarkers in Alzheimer's Disease"[1] (see URL in the footnote). We use 147 measures from the proteomic data downloaded from the ADNI web site.

The subjects of interest in AD research are divided into three broad categories: Control or normal cases (NC), mild cognitive impairment (MCI) cases and AD cases. The MCI cases, based on their status when followed-up over the course of a 4 year period, are further divided into MCI stable or non-converter cases (MCI NC), i.e., those MCI individuals who remain at MCI status and MCI converter cases (MCI C), i.e., those MCI patients who subsequently progress to AD. The summary of the number of samples for MRI and proteomics modalities, which passed the quality control and were available for the current study, and their baseline features together with disease status, are listed in Table 1. The data imbalance problem is clearly shown in Table 1. For example, in Table 1, AD cases are nearly double the number of control cases for the proteomics modality.

We examined both negative and positive class imbalances depending upon the prediction task and the feature set used. In proteomics measurements, there are 58 control cases (treated as negative class) versus 391 MCI cases (including both stable and converters; treated as positive class). For MRI modality, there are 191 control cases (treated as negative

[1] http://adni.loni.ucla.edu/wp-content/uploads/2010/11/BC_Plasma_Proteomics_Data_Primer.pdf

class) and 138 AD cases (treated as positive class). Disease prognosis is a critical task as the penalty attached to incorrect prediction is more than monetary. AD studies are targeted to provide early treatment to probable AD cases and to prevent or delay AD progression in AD cases. Incorrectly predicting an AD case as normal will prevent the patient from getting the required (or timely) medical treatment thereby reducing the patient's life expectancy. On the other hand, incorrect prediction of AD as a control case might cause distress to the patient and the family. Hence, it is challenging to determine the optimal costs to positive or negative class instances. Given the subtle and critical nature of the domain, in this study, we thoroughly examined different data re-sampling approaches and proposed a simple and versatile ensemble model approach to effectively handle class imbalance situation in the ADNI dataset.

2.2. Ensemble Framework

The ensemble system proposed in this study is a combination of data re-sampling technique, feature selection algorithm, and binary prediction model. The proposed ensemble system belongs to the class of external approaches with algorithmic level solutions. As noted earlier, external approaches for class imbalance problems are easily adaptable and are independent of the feature selection or classification algorithms. Furthermore, based on the domain requirements, algorithmic level solutions can be integrated with the proposed model to generate customized sophisticated learning model. This demonstrates the simplicity and versatility of our ensemble system. Within the proposed ensemble system, we analyze four basic data sampling approaches in addition to the no sampling approach, six feature selection algorithms, and two classification algorithms. The following are the data sampling approaches studied in this paper:

1. *No Sampling*: All of the data points from majority and minority training sets are used.
2. *Random Undersampling*: All of the training data points from the minority class are used. Instances are randomly removed from the majority training set till the desired balance is achieved. One disadvantage of this approach is that some useful information might be lost from the majority class due to the undersampling. This will be referred to as "Random US" in the following tables and figures.
3. *Random Oversampling*: All data points from majority and minority training sets are used. Additionally, instances are randomly picked, with replacement, from the minority training set till the desired balance is achieved. Adding the same minority samples might result in overfitting, thereby reducing the generalization ability of the classifier. This will be referred to as "Random OS" in the following tables and figures.
4. *K-Medoids Undersampling*: This is based on an unsupervised clustering algorithm in which the cluster centers are the actual data points. The majority training set is clustered where the number of clusters equals the number of minority training examples. Since, the initial cluster centers are chosen randomly, the process is repeated and the best result (the one with the minimum cost) is selected. The final training set is a combination of all data from the minority training set and the cluster centers from the majority training set. This approach is used only for undersampling, hence it will be referred as "K-Medoids" for the rest of this paper.
5. *SMOTE Oversampling:* SMOTE is the acronym for "Synthetic Minority Over-sampling Technique" which generates new synthetic data by randomly interpolating pairs of nearest neighbors. Details of the SMOTE algorithm can be found in the work by Chawla et al. (2002). This study used SMOTE to generate new synthetic data for the minority training set. The final training set is a

combination of all data from the majority and minority training sets and, additionally, the new synthetic minority data such that final training set is balanced. In this paper we use SMOTE only for oversampling, and it will be referred as "SMOTE" in the following figures and tables.

As noted earlier, an important goal of AD research is to identify key bio-signatures. The bio-signature discovery is done through feature selection which is defined as the process of finding a subset of relevant features (biomarkers) to develop efficient and robust learning models. Feature selection is an active research topic in the machine learning field. Based on prior work involving analysis of feature selection algorithms for bio-signature discovery in ADNI data (Dubey, 2012), this work explored the following six top-performing state-of-the-art feature selection algorithms: (1) two tailed Student's *t*-test [2] (referred to as T-Test); (2) Relief-F [3] based on relevance of features using k-nearest neighbors; (3) Gini Index[3] based on measure of inequality in the frequency distribution values; (4) Information Gain[3] which measures the reduction in uncertainty in predicting the class label; (5) Chi-Square[3] test for independence to determine whether the outcome is dependent on a feature; and (6) sparse logistic regression with stability selection (Meinshausen and Bühlmann, 2010) (referred to as SLR+SS) to select relevant features. A detailed description of feature selection algorithms can be found in Appendix.

In addition, two classifiers including Random Forest (RF) and Support Vector Machine (SVM) were used for classification using the top features selected. The framework for the ensemble system is illustrated in Figure 1. The graphical illustration of the basic data resampling techniques discussed above is shown in Figure 2 and Figure 3. Intuitively, one of the advantages of the undersampling over oversampling approach is that it reduces the overall training data size thereby saving memory and speeding up the classification process. In many empirical studies, undersampling has outperformed oversampling (Japkowicz, 2000a; Drummond and Holte, 2003). In addition to these basic re-sampling approaches, different rates of re-sampling and combination re-sampling approaches were also explored in our study.

2.3. Detailed Ensemble Procedure

The mathematical formulation of the problem statement and the solution is defined as follows:

Set of feature selection algorithms:

F = {T-Test, Relief-F, Gini Index, Information Gain, Chi-Square, SLR+SS}

Set of class-imbalance handling approaches:

S = {Different types and rates of data re-sampling techniques}

Set of classification algorithms:

C = {Random Forest, Support Vector Machine}

An ensemble system is defined as follows:

$E = \{f, s, c\}$, where $f \in F$, $s \in S$, and $c \in C$

For any set X, $|X|$ is defined as the cardinality of the set.

[2]Matlab's ttest2 function was used.
[3]We used the Feature Selection package in Zhao, et al. (2011)

Hence there were |F|×|S|×|C| ensemble systems studied in this paper for a given prediction task as illustrated in Figure 1. In this work, the experiments were designed such that we evaluated each ensemble system using k-fold cross validation. The training set in each cross fold was sampled multiple times to reduce the bias due to random dataset generation, thus producing multiple learning models. These models were combined using *majority voting*, where the final label of an instance is decided based on the majority votes received from all the models. In case of tie, the probability of the estimation given by the model is taken into consideration. For example, if 30 models (using the same re-sampling technique on the training set) are trained to estimate the labels of a test set and 20 models assign a test data point to class 1 whereas remaining 10 models assign it to class 2, then the final label of this particular test data point is taken as class 1. We also reported the averaged performance of all models and used it as the baseline for comparison.

2.4. Experimental Setup

The experiments conducted in this study were designed to maximally reduce the bias introduced due to randomness and to generate empirically comparable ensemble models. The pre-processed data was then divided into majority and minority sub-datasets. 10-fold cross validation was used such that each sub-dataset was partitioned into a fixed 9:1 train-test ratio. The train and test sets from the respective classes were combined to generate a training dataset and a testing dataset. Data resampling techniques were applied to the training dataset whereas for a given prediction task, the testing dataset was kept constant between different resampling techniques for a fair comparison. For example, for the task of discriminating control from AD cases, random undersampling and SMOTE oversampling techniques used the same test set for a given cross fold. This approach facilitates accurate comparison of the efficacy of different models (refer to Figure 1). Each cross-fold had multiple training sets for various resampling techniques (except for no-sampling approach, where each cross fold had just 1 dataset) wherein the test set remained the same and the training set varied based on the type of data re-sampling technique employed. In case of K-Medoids undersampling, the process of choosing the cluster center is repeated 10 times and the set of cluster centers which gives the minimum cost is selected. The SMOTE oversampling algorithm can have many variations in the choice of the new data point (synthetic data) lying on the line segment joining two nearest neighbors. In this paper, we used the basic approach which randomly chooses the synthetic data point on the line segment. The stability selection procedure used 1000 bootstrap runs and selected those prominent features. The classifiers with default settings were used for all experiments in this study. The predictions obtained from the ensemble model were compared with clinical diagnosis to evaluate the efficacy of the model. The probability of the prediction, obtained from the classifier for each test instance was recorded for later use. The efficacy of different ensemble systems was compared using various performance metrics including *accuracy*, *sensitivity*, *specificity*, and *area under the ROC curve* (AUC). These metrics are defined as follows:

$$\text{Accuracy}\ (\%) = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100$$
$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$
$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

where TP refers to the number of samples correctly identified as positive (*True Positive*), FP refers to the number of samples incorrectly identified as positive (*False Positive*), TN refers to the number of samples correctly identified as negative (*True Negative*), and FN refers to the number of samples incorrectly identified as negative (*False Negative*). Accuracy measures the percentage of correct classifications by the model. Sensitivity, also known as

recall rate or *True Positive Rate* (TPR), is the proportion of positive samples who are correctly identified as positive. Specificity is the proportion of negative samples who are correctly identified as negative. It is also known as *False Positive Rate* (FPR). AUC is computed by averaging the trapezoidal approximations for the curve created by TPR and FPR. Multiple classification models were generated for every cross fold, each of which provides a prediction, positive or negative, for the given class instance. Accuracy, sensitivity, specificity, and AUC are computed by utilizing the majority labels as discussed in Section 2.3.

3. RESULTS

This section provides the details of the comprehensive experiments performed and results obtained to compare efficacy of different ensemble systems. This study was focused on binary classification problem of identifying control, MCI, and AD cases from one another. Only MRI and proteomics modalities were studied as these are among the most easily available features in the AD domain. This section is divided into four subsections where each subsection compares the proposed ensemble framework with traditional and/or sophisticated solutions for the class imbalance problem. In Section 3.1, feature selection algorithms and basic data re-sampling approaches (refer to Section 2.2) were compared for different prediction tasks and modalities. Some researchers examined the use of combination approaches where different resampling techniques were combined to achieve a balanced training set (Chawla et al., 2002). In Section 3.2 we studied such an approach and compared it with our proposed model. On the other hand, some researchers have questioned the need of a balanced training set and essayed imbalanced training sets obtained by different rates of data sampling (Estabrooks et al., 2004); we examined the effect of rate of resampling in Section 3.3. Finally, in Section 3.4 we compared the proposed approach with the multi-classifier multi-learner approach (Chan and Stolfo, 1998).

In the following tables and figures, "(−)" is used to represent the negative class, whereas "(+)" is used to represent the positive class. "RF Avg" and "SVM Avg" represent averaged performance measures and "RF MajVote" and "SVM MajVote" represent majority voting performance measures using RF and SVM classifiers.

3.1. Comparing basic data resampling techniques

For the task of predicting NC from MCI cases using proteomics measurements, we used 5 basic data re-sampling techniques (refer to Section 2.2) and each approach used 6 feature selection algorithms and 2 different classifiers, thus generating 60 (= 5×6×2) ensemble systems. Each ensemble system used 10 fold cross-validation and 30 random datasets in each cross-fold except the no-sampling approach, yielding 300 (=10×30) classification models. The data distribution for no sampling, undersampling (random and K-Medoids), and oversampling (random and SMOTE) techniques is summarized in Table 2. To evaluate the six feature selection algorithms, we compared the performance of the top features obtained from each of these algorithms. A few selected comparison graphs are illustrated in Figure 4. All other data resampling techniques produced similar results (Dubey, 2012). As seen from this figure, the performance metric increases smoothly and stabilizes after selecting top 10–12 features; hence the results reported in this study are for top 10 features. Comparison of the 6 feature selection algorithms for top 10 features using SVM classifier (since SVM gave better classification measures than RF in most cases), is illustrated in Figure 5. The absolute difference between sensitivity and specificity (referred to as *Sensitivity Specificity gap*) is displayed for each feature selection algorithm, which illustrates the classifier's effectiveness in handling the class imbalance. A smaller gap between sensitivity and specificity is desirable. Clearly, SLR+SS outperformed other feature selection algorithms in all experiments; the overall performance of T-Test and GiniIndex was better than the remaining

ones. Since T-Test is very popular in the neuroimaging domain, this work reports its performance along with SLR+SS for all following experiments. The results are summarized in Table 3. Note that for the sake of brevity, we only report the most significant and illustrating results here.

From Figure 5 and Table 3, undersampling approaches, specifically K-Medoids, obtained better classification performance for imbalanced ADNI data. SLR+SS performed better in K-Medoids than random under-sampling whereas other feature selection algorithms showed similar or slightly better performance for random under-sampling. These results corroborate the efficacy of the ensemble system composed of SLR+SS feature selection algorithm, K-Medoids data re-sampling method, and SVM classifier. Also, majority voting results were better than the respective averaged performance measures.

Similar observations were made for the NC/MCI prediction task using MRI features. The summary of datasets used is provided in Table 4 and the classification results are given in Table 5. Table 6 and Table 7 represent data distribution and prediction performance, respectively, of the classical NC/AD prediction task using proteomics features. The data and the performance measures of NC/AD task using MRI features are summarized in Table 8 and Table 9, respectively. In this case, we encountered negative class majority. The task of predicting NC from MCI Converters & AD cases experiences a significant class-imbalance situation. Table 10 and Table 11 summarize the data details and performance measures for this task using proteomics features. The MRI counterparts of this task are given in Table 12 and Table 13. From these six classification tasks, we conclude that the K-Medoids undersampling approach dominated the overall efficacy of the ensemble system more than any other factor.

3.2. Comparison with a combination scheme

Chawla et al. (2002) proposed a combination scheme by mixing different rates of oversampling (using SMOTE) and random undersampling to reverse the initial bias of the learner towards the majority class in favor of the minority class. The training set was not always balanced with respect to two classes; the approach forced the learner to experience varying degrees of undersampling such that at some higher degree of undersampling the minority class had larger presence in the training set. We examined their combination scheme approach for NC/MCI prediction task using proteomics data. The training set was re-sampled (undersampled/oversampled) at 0%, 10%, 20%, … 100%. 0% re-sampling is equivalent to “No Sampling” and 100% re-sampling is known as complete sampling or full sampling. Hence, in 100% undersampling, the majority class is reduced to match the minority class count and 100% oversampling increases the minority samples in the training set to match the majority class count. The computation of the resampling rate is a slightly modified version of the resampling rate calculation proposed by Estabrooks et al. (Estabrooks et al., 2004). Mathematically, the gap between majority and minority count is divided by the desired number of resamplings and is referred to as diffCount in this study. We started resampling the data at 10%, in increments of 10% till 100% resampling is achieved, hence the difference between majority and minority count was divided by 10. In case of undersampling, the majority class count is reduced by a multiple of diffCount. Similarly, a multiple of diffCount is used to increment the minority count in oversampling case. For example, if there are 52 negative samples and 356 positive samples available for training, and we are resampling at 10% as explained earlier, then the diffCount = (356 – 52)/10 = 30.4. Therefore, 40% undersampling gives 234 ($\approx 356 - 4 \times 30.4$) majority class count and a 30% oversampling gives 143 ($\approx 52 + 3 \times 30.4$)minority samples in the training set. In our experimental setup, the training set was always balanced using different rates of K-Medoids undersampling and SMOTE oversampling. Hence if the majority class was 20% undersampled, then the minority class was 80% oversampled. The data used in different

sampling rates is summarized in Table 14 and the data distribution is illustrated in Figure 6. As before, 144 (=6×12×2) ensemble systems were generated using six feature selection algorithms, 12 resampling techniques, and RF and SVM classifiers. From the classification results, summarized in Table 15, it is evident that complete K-Medoids undersampling (referred to as S0_K100) performs better than other resampling rates. Also, SLR+SS and SVM gave superior learning models and majority voting was more effective than simple averaging. These results are compared in Figure 7.

3.3. Comparing different rates of data resampling

Estabrooks et al. (2004) proposed a multiple resampling method, to efficiently learn from imbalanced data. They experimented with independently varying rates of oversampling and undersampling. They generated 20 datasets, 10 each for oversampling and undersampling, by increasing the resampling rate in increments of 10% till 100% resampling is achieved. From the experiments conducted on various domains, they concluded that optimal resampling rate depends upon the resampling strategy and it varies from domain to domain. In this paper, we studied effects of varying rates of oversampling and undersampling on NC/MCI prediction task for proteomics features. The experimental setup consisted of 10 cross folds, each having 10 datasets and 9:1 train-test ratio in each dataset. Only one of the two resampling approaches is utilized for a particular rate of resampling. Hence, the training set was not balanced except in the event of complete oversampling and undersampling. We used diffCount measure, as explained in previous experiments, to achieve varying rates of resampling and examined 20 resampling techniques. Table 16, Table 17 and Figure 8 summarize the data distribution used in this experiment. The results of comparison of classification efficacy for independently varying rates of under and over sampling approaches are provided in Table 18. This dataset was dominated by positive class samples; hence high sensitivity and low specificity were expected. As noted earlier, the effectiveness of a classification model is inversely proportional to the sensitivity-specificity gap. We used this criterion and observed that in the ADNI data set, the gap decreased with increasing level of oversampling (SMOTE) till 40% SMOTE and started increasing again. Whereas, the gap gradually decreased with increasing degrees of undersampling (K-Medoids) and the best results were achieved at 100% K-Medoids with high sensitivity (0.89), good specificity (0.812), high AUC (0.97), and accuracy (88%). Only the complete K-Medoids undersampling approach increased the specificity by more than 51%. The results for majority performance metrics are illustrated in Figure 9.

3.4. Comparison with a multi-classifier learning approach

Chan and Stolfo (1998) proposed a multi-classifier meta-learning approach and concluded that the training class distribution affects the performance of the learned classifiers and the natural distribution can be different from the desired training distribution that maximizes performance. Their model ensured that none of the data points were discarded. They split the majority class into non-overlapping subsets such that each subset is roughly the size of minority class. A classifier was trained on each of these subsets and the minority training set. Later, these classifiers were stacked together to build a final ensemble classifier. In our study on ADNI data for NC/MCI prediction task using proteomics modality, we studied Chan and Stolfo's approach. We used 52 (−) minority training samples and 356 (+) majority training samples, which gives, roughly, 1:7 minority-majority class ratios. We generated 7 datasets utilizing 7 non-overlapping subsets from majority training set for a given minority training set. The data distribution is graphically depicted in Figure 10. Three data resampling techniques were examined, namely, random undersampling, K-Medoids, and Chan and Stolfo's approach. The 7 datasets created in each cross fold utilized the respective resampling approach keeping the testing set fixed between all three techniques for a given cross fold. We used a simple combination scheme where the classifier performance from all

7 classification models for a cross fold was either averaged or taken as a majority vote. The results displayed here are averaged over all 10 cross folds. The results are summarized in Table 19 and Figure 11. We can observe from these results that Chan and stolfo's approach gave better accuracy but did not remove the bias towards minority class resulting in comparatively poor AUC value and sensitivity-specificity gap. K-Medoids and Random undersampling were able to bridge the gap between sensitivity and specificity with 88% accuracy and 0.93 AUC. This further demonstrates the effectiveness of our simple ensemble system for handling the imbalanced data.

4. DISCUSSION

This paper has two major contributions. First, we introduced a robust yet simple framework to address imbalance problem in classification study. Secondly, by a comprehensive set of experiments we demonstrated the supremacy of K-Medoid undersampling approach over other basic data re-sampling techniques in the ADNI dataset. We used the approach of completely balancing the training set with respect to the two classes by utilizing only one type of data resampling technique. To the best of our knowledge, this is the first study to systematically investigate the data imbalance issue in the ADNI dataset. In this pilot work, we used MRI and proteomics modalities in ADNI to assess whether one can still achieve reasonably balanced classification results on an imbalanced dataset. We also implemented and applied several state-of-the-art imbalanced data processing methods, applied them to ADNI dataset and compared their performance with our proposed ensemble framework. Our discovery may provide guidance for future experimental design and statistical integration on large scale neuroimaging datasets. ADNI provides us an ideal testbed for the developed algorithms and tools as the data is so diverse and complex, and its universal availability. Moreover, it is also becoming a model for other large data collection projects, and clinical trials, so there will be a flood of data with similar complexities. We hope our work will increase the interest in this ubiquitous and important problem and other groups may consider using this approach to deal with the imbalance in the training dataset when performing future classification studies on imbalanced datasets.

In the study, six feature selection algorithms and five basic data resampling techniques were compared for different prediction tasks and modalities. It was concluded that undersampling, in particular K-Medoids, yields better learning models than other resampling approaches. "No sampling" approach gave the highest test accuracy, but the results were biased towards the majority class as the classifiers tend to minimize the misclassification costs by classifying all samples into the majority class. This results in a huge gap between sensitivity and specificity measures. Data re-sampling approaches performed better in the class imbalance scenario. Random oversampling tends to overfit the training data as the data points were duplicated, whereas random undersampling may lead to loss of vital information as data points were randomly removed. SMOTE and K-Medoids sampling methods use heuristics to select/eliminate the data points, hence their performance was superior compared with the corresponding random resampling techniques. Undersampling performed better than the oversampling approach for all prediction tasks. This could potentially be due to that in undersampling the data points selected in the training set accurately represented the original class distribution, and the bias introduced, if any, in selecting the data points from the majority class was minimized. On the other hand, oversampling approaches could disturb the data distribution within the class either by overfitting or generating synthetic data points which do not follow the original class distribution as we have very little information about the minority class. Also, the majority voting results were shown to be better than the respective averaged performance measures, which demonstrates the effectiveness of performing multiple undersampling.

To corroborate our findings, we extended our study to include a few other data re-sampling approaches proposed by different researchers. The first experiment performed in this series was the comparison of our ensemble framework with the combination scheme proposed by Chawla et al. (2002) for the ADNI dataset. In our experimental setup, we ensured balanced training sets with varying degrees of undersampling (using K-Medoids) and oversampling (using SMOTE) as noted in Section 3.2. The results support our ensemble system where complete K-Medoids undersampling outperformed all other resampling approaches. These findings suggest that the complexity of ADNI dataset makes it difficult to generate synthetic data points which fit the natural class distribution well. On the other hand, undersampling selects the data points from the original class distribution and hence has lesser impact, most of which is taken care by repeated application of K-Medoids.

In analysis of different rates of data resampling where training data need not be balanced, we made the same observation of the superior performance of the ensemble system using complete K-Medoids undersampling (Section 3.3). The decreasing performance of oversampling as amount of SMOTE is increased, which again indicates the failure of synthetic data generation techniques for ADNI. The increasing percentage of K-Medoids not only reduces the gap between sensitivity and specificity, but it also tries to eliminate/reduce the class bias due to the majority class, which is a desirable property. We further compared our approach with multi-classifier meta-learning approach proposed by Chan and Stolfo (1998). Their approach splits the majority class into non-overlapping subsets such that each subset is roughly the size of minority class, different from random undersampling and our K-Medoids undersampling. Our experiments on ADNI data showed that both random and K-Medoids undersampling approaches outperformed Chan's approach.

Comparison with pioneering disease diagnosis research in ADNI

We compared our ensemble system's performance with some of the earlier work done on ADNI dataset. As noted earlier, MRI features are very popular among researchers owing to their widespread availability and high discriminative power (Dickerson et al., 2001; Devanand et al., 2007). Seminal research by (Ray et al., 2007) laid the ground for blood based proteins as biomarkers for early AD diagnosis (Gomez Ravetti and Moscato, 2008; O'Bryant et al., 2011; Johnstone et al., 2012).

Early identification of potential AD cases before any cognitive decline symptoms are visible has been examined by several studies. Ray and colleagues used molecular tests to identify 18 signaling proteins which could discriminate between control and AD cases with nearly 90% accuracy (Ray et al., 2007). Gomez Ravetti and Moscato (2008) identified a 5-protein signature from Ray et al.'s 18-protein set which achieved 96% accuracy in predicting non-demented from AD cases. Johnstone et al. (2012) identified an 11 protein signature on ADNI dataset using a multivariate approach based on combinatorial optimization ((*α, β*)-*k Feature Set Selection*). They achieved 86% sensitivity and 65% specificity when assessed on the full set of control and AD samples (54 and 112). They also studied balanced approaches using 54 samples from both classes and demonstrated balanced sensitivity and specificity measures of 73.1%. Shen et al. (2011) proposed elastic net classifiers based on regularized logistic regression. Shen and his group utilized ADNI dataset with 146 proteomics features, 57 total control subjects, and 106 total AD cases and achieved best accuracy of 83.7% and an AUC of 89.9%. These results are very close to our observations where our ensemble system composed of SLR+SS and no sampling approach yielded best accuracy of 84.86%, 91.67% sensitivity, 72.5% specificity, and 91.25% AUC using top 10 features (See Table 7). In terms of a balanced dataset using the undersampling approach and top 10 features, we achieved best accuracy of 84.16%, 83.33% sensitivity and 85.83% specificity, and an AUC of 91.94%.

Shen et al. (2011) used reduced MRI features and a subset of control and AD subjects (54 and 106) from ADNI samples reporting 86.6% prediction accuracy. Yang et al. (Yang et al., 2011) proposed an independent component analysis (ICA) based method for studying the discriminative power of MRI features by coupling ICA with the SVM classifier. Their experiments on ADNI dataset resulted in highest accuracy of 76.9% with 74% sensitivity and 79.5% specificity for control vs AD (236 vs 202) prediction task on ADNI dataset. Our ensemble framework for MRI features performed significantly better giving 87.38% accuracy, 83.3% sensitivity and 90.18% specificity using K-Medoids sampling approach and SLR+SS feature selection algorithm.

An intermediate stage of AD progression is MCI when the patient starts depicting signs of cognitive decline but is not completely demented. An examination of control and prodromal AD cases can give valuable information about initial signs and factors responsible for memory impairment. There are many prior works on the automated disease diagnosis problem, that include partial least square based feature selection on MRI (Zhou et al., 2011), feature extraction methods based on MRI data (Cuingnet et al., 2011), and support vector machines to combine MRI, PET and CSF, etc. (Kohannim et al., 2010). In a recent work on ADNI dataset, Johnstone et al. (2012) achieved 93.5% sensitivity and 66.9% specificity for the prediction task of control vs MCI converters (54 vs 163) using their multivariate approach. With balanced training data using 54 samples for both categories, they reported 74.3% sensitivity and 79.3% specificity. Shen et al. (2011) studied control vs MCI (57 vs 110) ADNI subjects for proteomics modality and observed highest accuracy of 87.4% and 95.3% AUC. We applied our algorithm to predict control from MCI subjects (including both converters and non-converters). The ensemble system of K-Medoids with SLR+SS algorithm resulted in 87.63% accuracy, with 87.58% sensitivity, and 88.33% specificity for top 10 features. The data imbalance ratio was 7:1 in our case but we still managed to get >85% values for all performance metrics. This clearly demonstrates the validity and potential of our method.

Many researchers have explored control vs MCI classification using MRI features where MCI cases include both converters and non-converters (Fan et al., 2008; Davatzikos et al., 2010; Liu et al., 2011; Shen et al., 2011; Yang et al., 2011). Shen and others (Shen et al., 2011) observed 74.3% classification accuracy for control vs MCI (57 vs 110) prediction task on ADNI dataset using reduced MRI feature set. Yang et al.'s ICA method coupled with SVM classifier on ADNI dataset was able to discriminate control from MCI cases (236 vs 410) with highest accuracy of 72%, 71.3% sensitivity, and 68.6% specificity (Yang et al., 2011). Our proposed ensemble framework composed of K-Medoids and SLR+SS gave 69.46% accuracy, 64% sensitivity, 79.5% specificity, and 77.15% AUC for the same prediction task.

In summary, although a direct head-to-head comparison is difficult (e.g. even the MRI features are different between studies), our experimental results were comparable or outperformed those of some state-of-the-art algorithms, e.g. (Cuingnet et al., 2011). More importantly, since we address a fundamental problem, we believe our work could be complementary to these existing research efforts and may help others to achieve a balanced and improved performance on the ADNI or other biomedical datasets.

5. CONCLUSION

Here we present a novel study in which different sampling approaches were thoroughly analyzed to determine their effectiveness in handling imbalanced neuroimaging data. This work demonstrates the efficacy of undersampling approach for class imbalance problem in ADNI dataset. In this work, several simple and robust ensemble systems were built based on

different data sampling approaches. Each ensemble system was composed of a feature selection algorithm and a data level solution for class imbalance problems (i.e. data resampling approach). We studied six state-of-the-art feature selection algorithms, namely, two tailed Student's *t*-test, Relief-F based on relevance of features using k-nearest neighbors, Gini Index based on measure of inequality in the frequency distribution values, Information Gain which measures the reduction in uncertainty in predicting the class label, Chi-Square test for independence to determine whether the outcome is dependent on a feature, and sparse logistic regression with stability selection. The data level resampling solutions studied in this work included random undersampling, random oversampling, K-Medoids based undersampling, and Synthetic Minority Oversampling Technique. We also experimented with different rates of under and over sampling and examined a combination data resampling approach where different rates of under and over sampling were combined together. The classification model was built using decision tree based Random Forest algorithm and decision boundary based Support Vector Machine classifiers. The key evaluation criteria used were accuracy and AUC curve along with sensitivity and specificity values. Since most resampling approaches randomly select the data points to remove or duplicate, the process was repeated a couple of times to remove any bias due to random selection. We compared the classification metrics obtained using averaged results and majority voting for all repetitions. The experiments conducted as a part of this study demonstrated the dominance of undersampling approaches over oversampling techniques. In general, sophisticated techniques such as K-Medoids and SMOTE gave better AUC and balanced sensitivity and specificity measures than the corresponding random resampling methods. This paper concludes that the ensemble system consisting of sparse logistic regression with stability selection as feature selection algorithm and K-Medoids complete undersampling approach (balanced train set with respect to the two classes) elegantly handles class imbalance problem in case of ADNI dataset. Performance metric based on majority voting dominates the corresponding averaged metric.

A concerted effort is needed to investigate the class imbalance problem in ADNI dataset. To the best of our knowledge, this is the first effort in that direction. This work studied proteomics and MRI modalities; future work will involve other MRI data features such as detailed tensor-based morphometry (TBM) features that were used in our voxelwise genome-wide association study (Stein et al., 2010a; Stein et al., 2010b; Hibar et al., 2011) and our surface multivariate TBM studies (Wang et al., 2011). Other modalities would also be considered, such as genomics, psychometric assessment scores, and clinical data. An integrative approach which uses a combination of different modalities can also be studied. Additionally, experiments can be performed on Alzheimer's disease datasets from other sources to check for common patterns.

In this study, we investigate feature selection for imbalanced data. Another popular approach for dimensionality reduction is feature extraction, e.g., principal component analysis or independent component analysis, which transforms the data into a different domain. The presented ensemble system can be extended to perform feature extraction and classification for imbalanced data. The current study focuses on binary classification. An interesting future direction is to extend the sampling techniques to the case of predictive regression (e.g., prediction of clinical measures). In this case, the distribution of the clinical measure should be taken into account when resampling the data. To the best of our knowledge, data resampling for regression has not been well studied in the literature. We plan to explore this in our future work.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

This work was funded by the National Institute on Aging (AG016570 to PMT and R21AG043760 to YW), the National Library of Medicine, the National Institute for Biomedical Imaging and Bioengineering, and the National Center for Research Resources (LM05639, EB01651, RR019771 to PMT), US National Science Foundation (NSF) (IIS-0812551, IIS-0953662 to JY), and National Library of Medicine (R01 LM010730 to JY).

References

Akbani, R.; Kwek, S.; Japkowicz, N. Applying support vector machines to imbalanced datasets. Proceedings of the 15th European Conference on Machine Learning (ECML); 2004. p. 39-50.

Alzheimer's Association. Alzheimer's Disease Facts and Figures. Alzheimer's Association; 2012. Available from: http://www.alz.org

Bartzokis G. Age-related myelin breakdown: a developmental model of cognitive decline and Alzheimer's disease. Neurobiol Aging. 2004; 25(1):5–18. author reply 49–62. [PubMed: 14675724]

Bernal-Rusiel JL, Greve DN, Reuter M, Fischl B, Sabuncu MR. Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models. Neuroimage. 2012; 66C:249–260. [PubMed: 23123680]

Bradford, JP.; Kunz, C.; Kohavi, R.; Brunk, C.; Brodley, CE. Pruning decision trees with misclassification costs. Proceedings of the European Conference on Machine Learning; 1998. p. 131-136.

Chan, PK.; Stolfo, SJ. Toward Scalable Learning with Non-uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining; AAAI Press; 1998. p. 164-168.

Chawla, N.; Japkowicz, N.; Ko cz, A. ICML'2003 Workshop on Learning from Imbalanced Data Sets (II); Washington DC, US. 2003.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Int Res. 2002; 16(1):321–357.

Chawla NV, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. SIGKDD Explor Newsl. 2004; 6(1):1–6.

Chen, C.; Liaw, A.; Breiman, L. Using Random Forest to Learn Imbalanced Data. University of California; Berkeley: 2004.

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 1993; 261(5123):921–923. [PubMed: 8346443]

Cover, TM.; Thomas, JA. Elements of Information Theory. Wiley; 1991.

Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehericy S, Habert MO, Chupin M, Benali H, Colliot O. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. Neuroimage. 2011; 56(2)

Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol Aging. 2010

Devanand DP, Pradhaban G, Liu X, Khandji A, De Santi S, Segal S, Rusinek H, Pelton GH, Honig LS, Mayeux R, Stern Y, Tabert MH, de Leon MJ. Hippocampal and entorhinal atrophy in mild cognitive impairment: prediction of Alzheimer disease. Neurology. 2007; 68(11):828–836. [PubMed: 17353470]

Dickerson BC, Goncharova I, Sullivan MP, Forchetti C, Wilson RS, Bennett DA, Beckett LA, deToledo-Morrell L. MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer's disease. Neurobiol Aging. 2001; 22(5):747–754. [PubMed: 11705634]

Drummond, C.; Holte, RC. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. Working Notes of the ICML'03 Workshop on Learning from Imbalanced Data Sets; Washington, DC. 2003.

Dubey, R. Masters Thesis. Arizona State University; 2012. Machine Learning Methods for Biosignature Discovery.

Duchesnay E, Cachia A, Boddaert N, Chabane N, Mangin JF, Martinot JL, Brunelle F, Zilbovicius M. Feature selection and classification of imbalanced datasets: application to PET images of children with autistic spectrum disorders. Neuroimage. 2011; 57(3):1003–1014. [PubMed: 21600290]

Duchi, J.; Shalev-Shwartz, S.; Singer, Y.; Chandra, T. Efficient projections onto the l_1-ball for learning in high dimensions. Proceedings of the 25th international conference on Machine learning; Helsinki, Finland: ACM; 2008. p. 272-279.

Elkan, C. The foundations of cost-sensitive learning. Proceedings of the 17th international joint conference on Artificial intelligence; Seattle, WA, USA: Morgan Kaufmann Publishers Inc; 2001. p. 973-978.

Elkan, C. Invited talk: The real challenges in data mining: A contrarian view. 2003. http://www.site.uottawa.ca/~nat/Workshop2003/realchallenges2.ppt

Ertekin, S.; Huang, J.; Bottou, L.; Giles, L. Learning on the border: active learning in imbalanced data classification. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management; Lisbon, Portugal: ACM; 2007. p. 127-136.

Estabrooks, A. Master thesis. Computer Science, Dalhousie University; 2000. A combination scheme for inductive learning from imbalanced data sets.

Estabrooks A, Jo T, Japkowicz N. A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence. 2004; 20(1):18–36.

Fan Y, Resnick SM, Wu X, Davatzikos C. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. NeuroImage. 2008; 41(2): 277–285. [PubMed: 18400519]

Fitzmaurice, G.; Laird, N.; Ware, J. Applied longitudinal analysis. Wiley; 2011.

Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010; 33(1):1–22. [PubMed: 20808728]

Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol. 2010; 6(2):67–77. [PubMed: 20139996]

Fu W. Penalized Regressions: The Bridge versus the Lasso. Journal of Computational and Graphical Statistics. 1998; 7(3):397–416.

Gomez Ravetti M, Moscato P. Identification of a 5-protein biomarker molecular signature for predicting Alzheimer's disease. PLoS One. 2008; 3(9):e3111. [PubMed: 18769539]

He H, Garcia EA. Learning from Imbalanced Data. Knowledge and Data Engineering, IEEE Transactions on. 2009; 21(9):1263–1284.

Hibar DP, Stein JL, Kohannim O, Jahanshad N, Saykin AJ, Shen L, Kim S, Pankratz N, Foroud T, Huentelman MJ, Potkin SG, Jack CR Jr, Weiner MW, Toga AW, Thompson PM. Voxelwise gene-wide association study (vGeneWAS): Multivariate gene-based association testing in 731 elderly subjects. Neuroimage. 2011; 56(4):1875–1891. [PubMed: 21497199]

Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, Whitwell JL, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DLG, Killiany R, Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS, Krueger G, Ward HA, Metzger GJ, Scott KT, Mallozzi R, Blezek D, Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW, Study A. The Alzheimer's disease neuroimaging initiative (ADNI):

MRI methods. Journal of Magnetic Resonance Imaging. 2008; 27(4):685–691. [PubMed: 18302232]

Japkowicz, N. In: Japkowicz, N., editor. Learning from Imbalanced Data Sets: A Comparison of Various Strategies; Proceedings of Learning from Imbalanced Data Sets, Papers from the AAAI Workshop; 2000a. p. 10-15.

Japkowicz, N. The Class Imbalance Problem: Significance and Strategies. Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI); 2000b. p. 111-117.

Japkowicz N. Supervised Versus Unsupervised Binary-Learning by Feedforward Neural Networks. Mach Learn. 2001; 42(1–2):97–122.

Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intell Data Anal. 2002; 6(5):429–449.

Jiang X, El-Kareh R, Ohno-Machado L. Improving predictions in imbalanced data using Pairwise Expanded Logistic Regression. AMIA Annu Symp Proc. 2011; 2011:625–634. [PubMed: 22195118]

Jo T, Japkowicz N. Class imbalances versus small disjuncts. SIGKDD Explor Newsl. 2004; 6(1):40–49.

Johnstone D, Milward EA, Berretta R, Moscato P. Multivariate protein signatures of pre-clinical Alzheimer's disease in the Alzheimer's disease neuroimaging initiative (ADNI) plasma proteome dataset. PLoS One. 2012; 7(4):e34341. [PubMed: 22485168]

Joshi, MV.; Kumar, V.; Agarwal, RC. Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements. Proceedings of the 2001 IEEE International Conference on Data Mining. IEEE Computer Society; 2001. p. 257-264.

Knoll U, Nakhaeizadeh G, Tausend B. Cost-sensitive pruning of decision trees. Machine Learning: ECML-94. 1994; 784:383–386.

Kohannim O, Hua X, Hibar DP, Lee S, Chou YY, Toga AW, Jack CR Jr, Weiner MW, Thompson PM. Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiol Aging. 2010; 31(8):1429–1442. [PubMed: 20541286]

Ko cz, A.; Chowdhury, A.; Alspector, J. Data duplication: An imbalance problem?. Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets; 2003.

Kubat, M.; Matwin, S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. Proceedings of the Fourteenth International Conference on Machine Learning; Morgan Kaufmann; 1997. p. 179-186.

Lee KJ, Hwang YS, Kim S, Rim HC. Biomedical named entity recognition using two-phase model based on SVMs. J Biomed Inform. 2004; 37(6):436–447. [PubMed: 15542017]

Ling, C.; Li, C. Data Mining for Direct Marketing: Problems and Solutions. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98); AAAI Press; 1998. p. 73-79.

Liu, J.; Chen, J.; Ye, J. Large-scale sparse logistic regression. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining; Paris, France: ACM; 2009a. p. 547-556.

Liu, J.; Ji, S.; Ye, J. SLEP: Sparse Learning with Efficient Projections. Arizona State University; 2009b. http://www.public.asu.edu/~jye02/Software/SLEP

Liu XY, Wu J, Zhou ZH. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern B Cybern. 2009c; 39(2):539–550. [PubMed: 19095540]

Liu Y, Paajanen T, Zhang Y, Westman E, Wahlund LO, Simmons A, Tunnard C, Sobow T, Mecocci P, Tsolaki M, Vellas B, Muehlboeck S, Evans A, Spenger C, Lovestone S, Soininen H. Combination analysis of neuropsychological tests and structural MRI measures in differentiating AD, MCI and control groups--the AddNeuroMed study. Neurobiol Aging. 2011; 32(7):1198–1206. [PubMed: 19683363]

Maloof, MA. Learning when data sets are imbalanced and when costs are unequal and unknown. ICML-2003 Workshop on Learning from Imbalanced Data Sets II; 2003.

Mayeux R, Saunders AM, Shea S, Mirra S, Evans D, Roses AD, Hyman BT, Crain B, Tang MX, Phelps CH. Utility of the apolipoprotein E genotype in the diagnosis of Alzheimer's disease.

Alzheimer's Disease Centers Consortium on Apolipoprotein E and Alzheimer's Disease. N Engl J Med. 1998; 338(8):506–511. [PubMed: 9468467]

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

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L. 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. 2005; 1(1):55–66.

O'Bryant SE, Xiao G, Barber R, Huebinger R, Wilhelmsen K, Edwards M, Graff-Radford N, Doody R, Diaz-Arrastia R. A blood-based screening tool for Alzheimer's disease that spans serum and plasma: findings from TARC and ADNI. PLoS One. 2011; 6(12):e28092. [PubMed: 22163278]

Padmaja, TM.; Krishna, PR.; Bapi, RS. Majority filter-based minority prediction (MFMP): An approach for unbalanced datasets. TENCON 2008 – 2008 IEEE Region 10 Conference; 2008. p. 1-6.

Pazzani, M.; Merz, C.; Murphy, P.; Ali, K.; Hume, T.; Brunk, C. Reducing misclassification costs. Proceedings of the 11th International Conference on Machine Learning; 1994. p. 217-225.

Provost, F. Machine Learning from Imbalanced Data Sets 101. Workshop on Learning from Imbalanced Data Sets; Texas, US: AAAI; 2000.

Provost F, Fawcett T. Robust Classification for Imprecise Environments. Mach Learn. 2001; 42(3): 203–231.

Ray S, Britschgi M, Herbert C, Takeda-Uchimura Y, Boxer A, Blennow K, Friedman LF, Galasko DR, Jutel M, Karydas A, Kaye JA, Leszek J, Miller BL, Minthon L, Quinn JF, Rabinovici GD, Robinson WH, Sabbagh MN, So YT, Sparks DL, Tabaton M, Tinklenberg J, Yesavage JA, Tibshirani R, Wyss-Coray T. Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. Nat Med. 2007; 13(11):1359–1362. [PubMed: 17934472]

Reiman EM, Jagust WJ. Brain imaging in the study of Alzheimer's disease. Neuroimage. 2011

Robnik-ikonja M, Kononenko I. Theoretical and Empirical Analysis of ReliefF and RReliefF. Mach Learn. 2003; 53(1–2):23–69.

Shaw LM, Vanderstichele H, Knapik-Czajka M, Clark CM, Aisen PS, Petersen RC, Blennow K, Soares H, Simon A, Lewczuk P, Dean R, Siemers E, Potter W, Lee VM, Trojanowski JQ. Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. Ann Neurol. 2009; 65(4):403–413. [PubMed: 19296504]

Shen, L.; Kim, S.; Qi, Y.; Inlow, M.; Swaminathan, S.; Nho, K.; Wan, J.; Risacher, SL.; Shaw, LM.; Trojanowski, JQ.; Weiner, MW.; Saykin, AJ. Identifying neuroimaging and proteomic biomarkers for MCI and AD via the elastic net. Proceedings of the First international conference on Multimodal brain image analysis; Toronto, Canada: Springer-Verlag; 2011. p. 27-34.

Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR Jr, Kaye J, Montine TJ, Park DC, Reiman EM, Rowe CC, Siemers E, Stern Y, Yaffe K, Carrillo MC, Thies B, Morrison-Bogorad M, Wagster MV, Phelps CH. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011; 7(3):280–292. [PubMed: 21514248]

Stein JL, Hua X, Lee S, Ho AJ, Leow AD, Toga AW, Saykin AJ, Shen L, Foroud T, Pankratz N, Huentelman MJ, Craig DW, Gerber JD, Allen AN, Corneveaux JJ, Dechairo BM, Potkin SG, Weiner MW, Thompson PM. Voxelwise genome-wide association study (vGWAS). Neuroimage. 2010a; 53(3):1160–1174. [PubMed: 20171287]

Stein JL, Hua X, Morra JH, Lee S, Hibar DP, Ho AJ, Leow AD, Toga AW, Sul JH, Kang HM, Eskin E, Saykin AJ, Shen L, Foroud T, Pankratz N, Huentelman MJ, Craig DW, Gerber JD, Allen AN, Corneveaux JJ, Stephan DA, Webster J, DeChairo BM, Potkin SG, Jack CR Jr, Weiner MW, Thompson PM. Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease. Neuroimage. 2010b; 51(2):542–554. [PubMed: 20197096]

Van Hulse, J.; Khoshgoftaar, TM.; Napolitano, A. Experimental perspectives on learning from imbalanced data. Proceedings of the 24th international conference on Machine learning; Corvalis, Oregon: ACM; 2007. p. 935-942.

Visa, S.; Ralescu, A. Issues in mining imbalanced data sets - a review paper. Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference; 2005. p. 67-73.

Vlkolinsk R, Cairns N, Fountoulakis M, Lubec G. Decreased brain levels of 2′,3′-cyclic nucleotide-3′-phosphodiesterase in Down syndrome and Alzheimer's disease. Neurobiol Aging. 2001; 22(4): 547–553. [PubMed: 11445254]

Wang Y, Song Y, Rajagopalan P, An T, Liu K, Chou YY, Gutman B, Toga AW, Thompson PM. Surface-based TBM boosts power to detect disease effects on the brain: An N=804 ADNI study. Neuroimage. 2011; 56(4):1993–2010. [PubMed: 21440071]

Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, Harvey D, Jack CR, Jagust W, Liu E, Morris JC, Petersen RC, Saykin AJ, Schmidt ME, Shaw L, Siuciak JA, Soares H, Toga AW, Trojanowski JQ. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimers Dement. 2012; 8(1 Suppl):S1–68. [PubMed: 22047634]

Yang Q, Wu X. 10 Challenging Problems in Data Mining Research. International Journal of Information Technology & Decision Making. 2006; 5(4):597–604.

Yang W, Lui RL, Gao JH, Chan TF, Yau ST, Sperling RA, Huang X. Independent component analysis-based classification of Alzheimer's disease MRI data. J Alzheimers Dis. 2011; 24(4): 775–783. [PubMed: 21321398]

Yen, S-J.; Lee, Y-S. Cluster-Based Sampling Approaches to Imbalanced Data Distributions. In: Tjoa, A.; Trujillo, J., editors. Data Warehousing and Knowledge Discovery. Springer; Berlin Heidelberg: 2006. p. 427-436.

Yuan L, Wang Y, Thompson PM, Narayan VA, Ye J. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. Neuroimage. 2012; 61(3):622–632. [PubMed: 22498655]

Zadrozny, B.; Langford, J.; Abe, N. Cost-Sensitive Learning by Cost-Proportionate Example Weighting. Proceedings of the Third IEEE International Conference on Data Mining; IEEE Computer Society; 2003. p. 435

Zhao, Z.; Morstatter, F.; Sharma, S.; Alelyani, S.; Anand, A.; Liu, H. ASU Feature Selection Repository. 2011. Advancing feature selection research.

Zheng, Z.; Srihari, R. Optimally combining positive and negative features for text categorization. Workshop for Learning from Imbalanced Datasets II, Proceedings of the (ICML); 2003.

Zhou L, Wang Y, Li Y, Yap PT, Shen D. Hierarchical anatomical brain networks for MCI prediction: revisiting volumetric measures. PLoS One. 2011; 6(7):e21935. [PubMed: 21818280]

APPENDIX

In this appendix, we detail the six feature selection algorithms which were adopted in our experiments.

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, symmetric to normal distribution but with unknown variance. This work employed unpaired two-tailed *t*-test which 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 variance. The *p*-value is computed for each feature independently using *t*-score (test statistics) and is defined as the probability of observing a sample statistic as extreme or more extreme as test statistic under the null hypothesis. The null hypothesis is rejected if *p*-value is less than or equal to the

significance level, usually denoted by $\alpha \leq 0.05$. Features are arranged in increasing order of *p*-value such that the most important feature has least *p*-value. The matlab's builtin T-Test function is used for this algorithm.

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. The underlying principle states that for each feature, if the distance between data points from the same classes is large, then the feature distinguishes data points within the same class. Such a feature is of no use and hence its weight should be reduced. Whereas if the difference between data points from different classes is large, then the feature distinguishes the data points from two different classes which serves the feature selection problem formulation well. The weights of such features are increased. Thus, the significant features are arranged in descending order of their weights. The Relief-F algorithm improves the Relief algorithm by introducing k-nearest neighbors from each class (Robnik-ikonja and Kononenko, 2003).

Gini Index

Gini Index (GI), also known as Gini Coefficient or Gini Ratio, measures the inequality in the frequency distribution values. This statistical measure of dispersion is commonly used to measure wealth or income inequality within the population or among countries. It can be applied to various other fields as well. Mathematically it is defined as the ratio of the area within the Lorenz curve and the line of equality [18]. GI measures the ability of a feature to differentiate between target classes. When all the samples belong to the same target class, GI is zero indicating maximum inequality thereby giving most useful information. On the other hand, if all samples are equally distributed between target classes, then GI reaches its maximum value denoting minimum information which can be obtained from this feature. Hence, features are arranged in increasing order of GI where most significant feature has least GI.

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 (Cover and Thomas, 1991). It represents the reduction in uncertainty of predicting class label (Y) given a feature vector (x_a) which can take up to k possible values. In other words, IG measures the reduction in entropy in moving from a prior distribution $P(Y)$ to a posterior distribution $P(Y|x_a)$. Both Y and x_a are assumed to be discrete. An attribute with higher value of IG is considered to be more relevant and is assigned a higher weight. Features are arranged in decreasing order of their IG values. This is an asymmetric method, i.e. $IG(Y|x_a) \neq IG(x_a|Y)$, and is not suitable for attributes (feature vectors) which can take a large number of discrete values as it might cause overfitting problems.

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. It is a continuous, asymmetrical, skewed to right distribution, and has K degrees of freedom such that the mean of the distribution is equal to

and K the variance is $2K$. 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 of the outcomes of an observation are statistically independent. P-value is the probability of obtaining a test statistic as extreme as the observed value under null hypothesis and is computed from distribution χ^2 table given χ^2 test statistic and K. The null hypothesis is rejected if p-value is less than the specified significance level α, which is often $\alpha \leq 0.05$. Rejecting the hypothesis makes the result statistically significant and confirms the dependence of the outcome on the feature value. Features are arranged in increasing order of p-value.

Sparse Logistic Regression

Sparse Logistic Regression (SLR) is an embedded feature selection algorithm which uses ℓ_1-norm regularization in Logistic Regression. It is one of the most attractive feature selection algorithms in applications which deal with high dimensional data. Logistic Regression (LR) 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 sigmoid function ($h_\theta(x)$) such that if $h_\theta(x) \geq 0.5$, the class label is positive otherwise it is negative. LR tends to overfit when the sample size is limited and the data is very high dimensional. To reduce overfitting and obtain better LR classifiers, regularization is applied to the LR's objective function. The guiding principle in sparse logistic regression is to use regularization in Logistic loss function such that irrelevant features are given a zero weight (Liu et al., 2009a). To induce sparsity, ℓ_1-norm regularized logistic loss function is used (Fu, 1998; Duchi et al., 2008; Friedman et al., 2010). Features are ranked in decreasing order of their weights. The matlab code is taken from the SLEP package (Liu et al., 2009b).

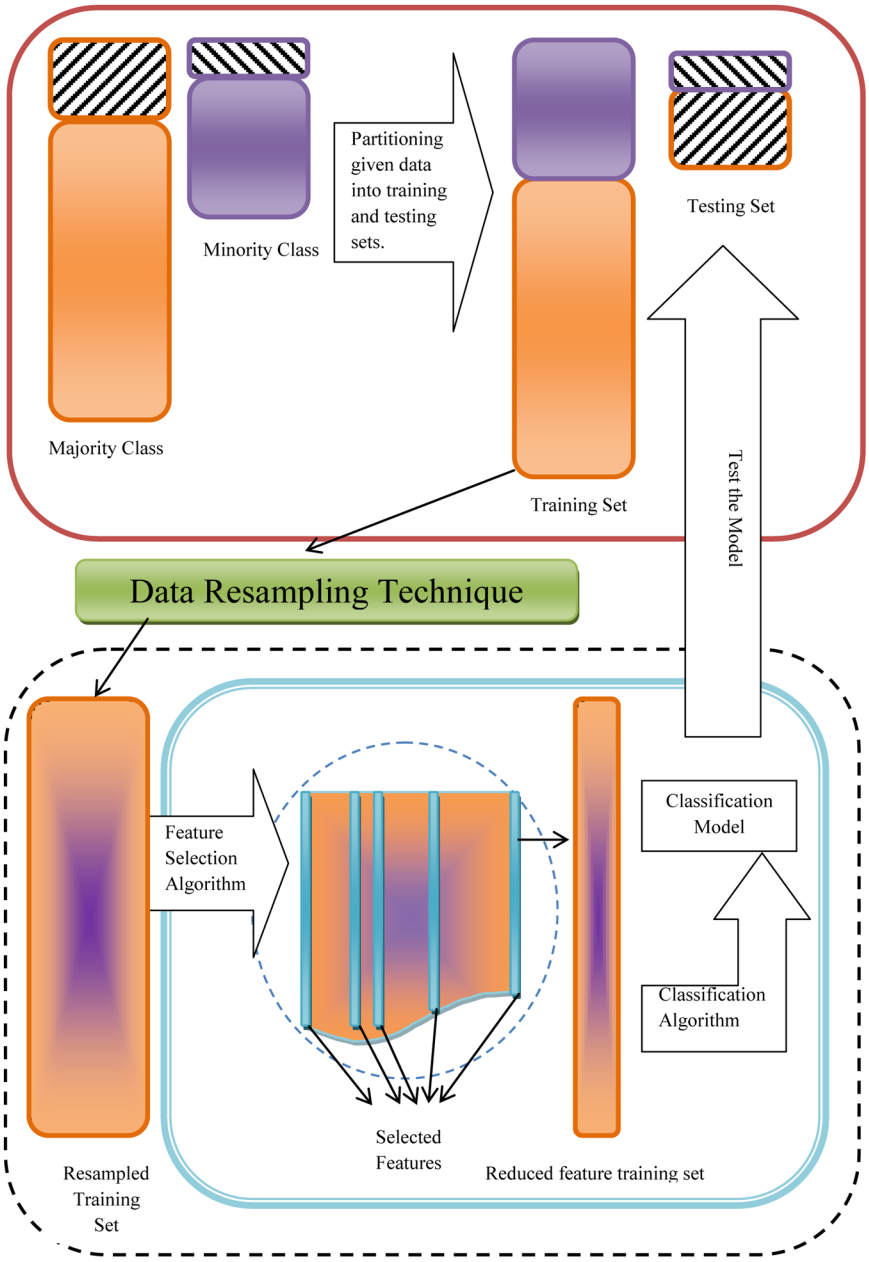

Figure 1.

Illustrating the proposed ensemble system for imbalance data classification. In this proposed model, a training and a testing set is derived from the given data using data points from both majority and minority classes as illustrated in the top rectangle (solid line) of the figure. Different data re-sampling techniques are applied to the training set to generate a "re-sampled training set" on which a feature selection algorithm is applied to select relevant features resulting in a reduced dimension training set. Subsequently a classification algorithm is applied to generate a prediction model which is tested on the test set to evaluate its efficacy. The steps shown in double blue bordered rectangle are repeated for each feature selection algorithm and prediction model. The steps in dotted black bordered rectangle are repeated for each data resampling technique.

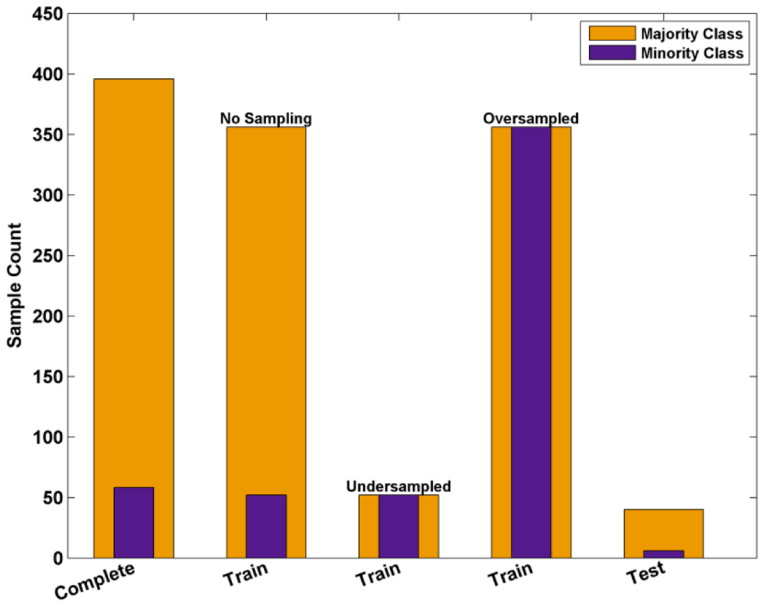

Figure 2.

This example illustrates class imbalance problem and the basic data resampling techniques used in the ADNI dataset for predicting MCI from Control cases on proteomics features (refer Table 1). The bar labeled “Complete” represents the data available for analysis. The “Train” bar represents training data taken from both classes for different resampling approaches and “Test” bar represents the test data. A dataset is formed by combining a training set and a test set (test set is kept fixed between different sampling approaches, and it need not be balanced).

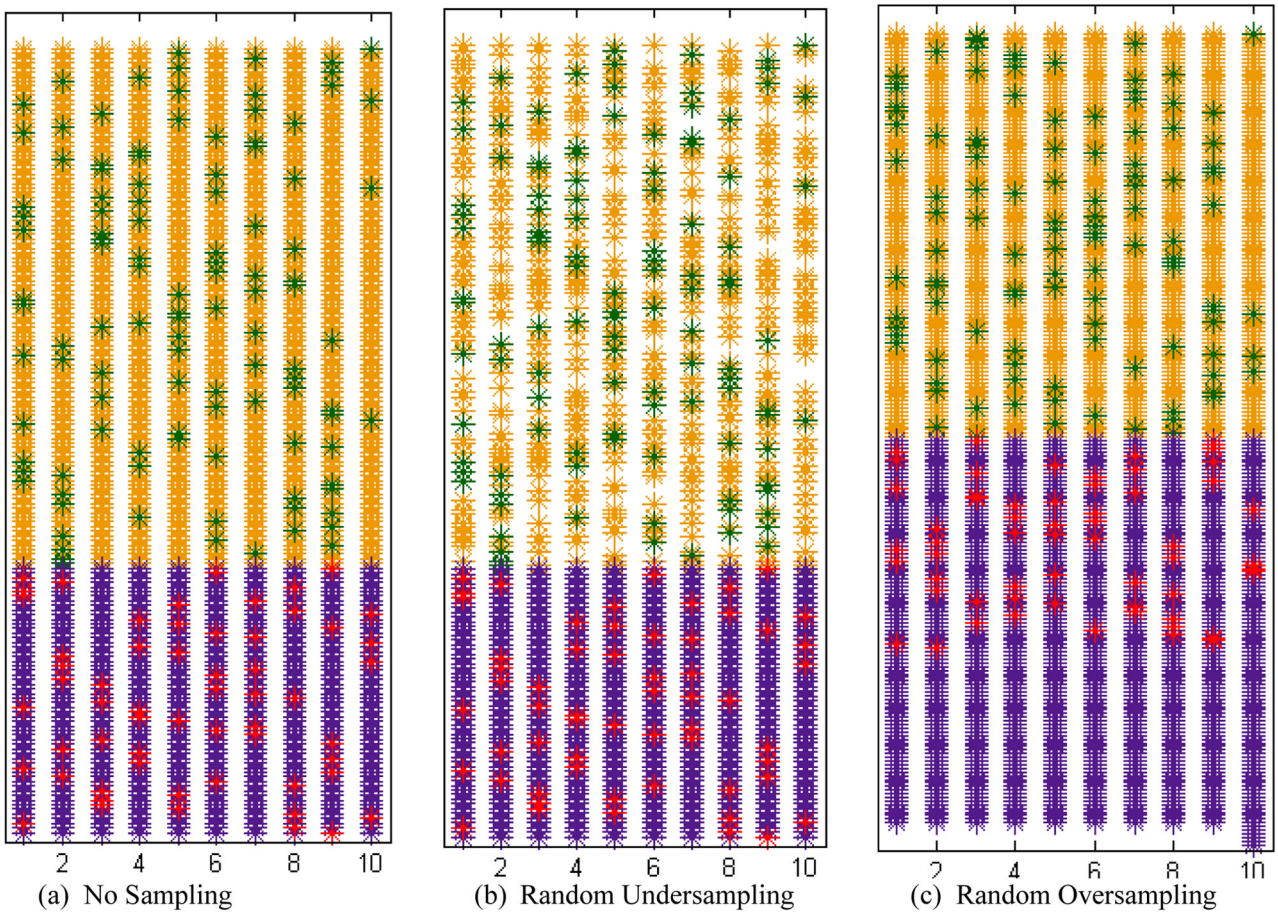

Figure 3.

Illustrating three different sampling approaches used in an ensemble system for an experimental setup for predicting control cases (marked by blue, for training and red, for testing asterisk symbols) from AD cases (marked by orange, for training and green, for testing asterisk symbols) using proteomics modality (refer to Table 1). Each class is divided into a training and test set in a ratio of 9:1. X-axis represents 10 cross folds and Y-axis represents samples. Fig. (a) depicts actual or no sampling scenario where training data is unbalanced with respect to the two classes. Fig. (b) depicts undersampling scenario where training set is balanced by removing data points from the majority class as shown by the sparse orange columns for each cross fold compared to other two cases. Fig (c) depicts oversampling scenario where minority class is duplicated as shown by extra length of blue columns for each cross fold. Note that only one dataset is shown for each cross fold, but 30 datasets were used except in training for no sampling case.

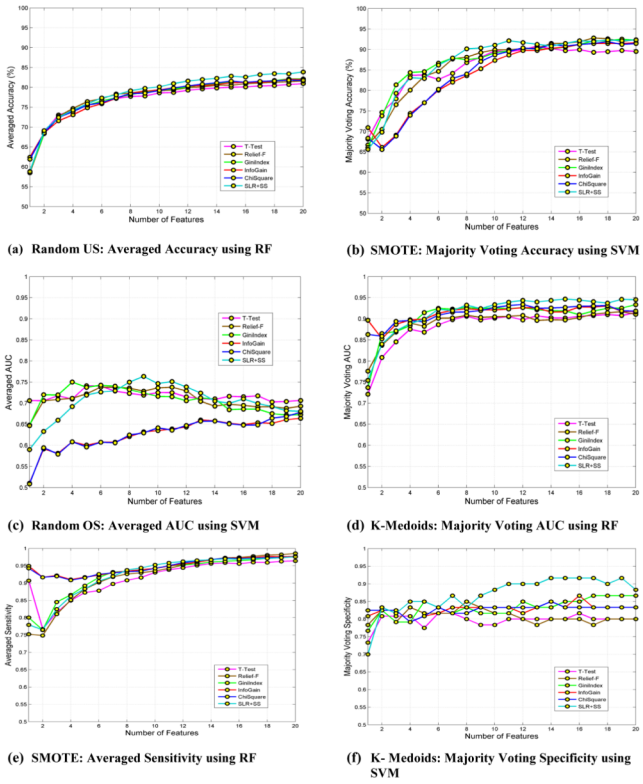

Figure 4.

NC/MCI prediction task: Comparison of feature selection algorithms for different performance metrics, classifiers, and sampling approaches. The results were averaged across 10 cross folds for top 20 features.

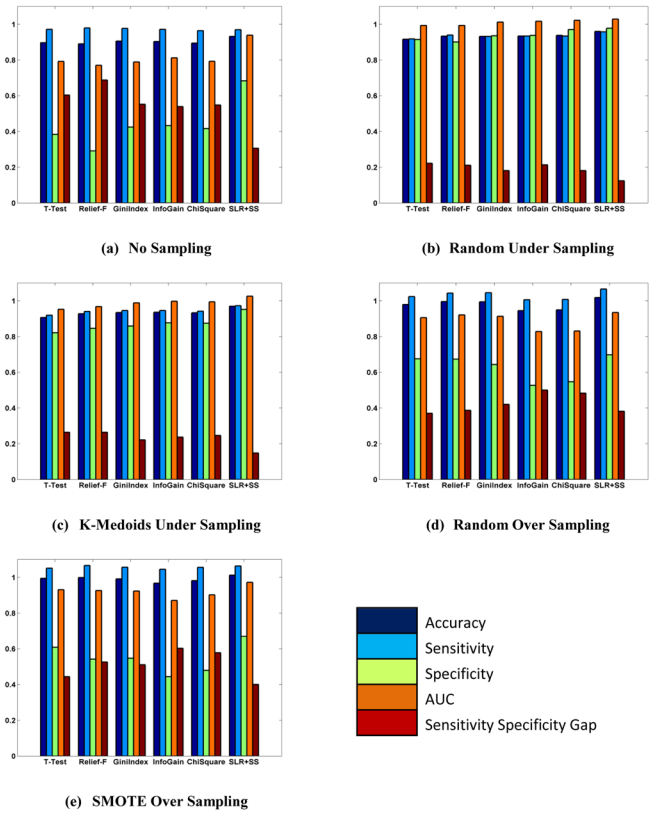

Figure 5.
NC/MCI majority voting classification performance comparison of SVM classifier, averaged across 10 cross folds, using top 10 features from six feature selection algorithms for different data sampling approaches.

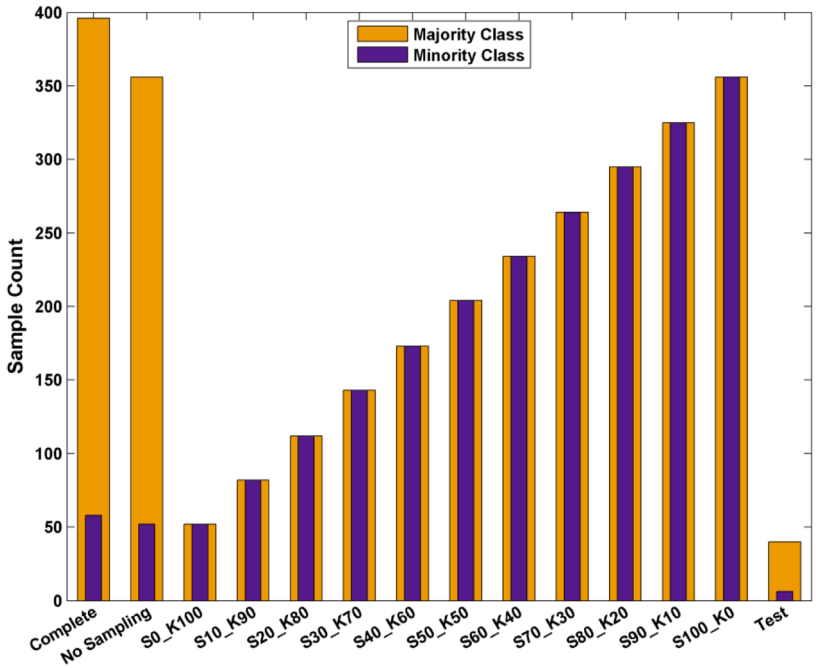

Figure 6.

The bar labeled “Complete” represents the data available for analysis. The “Test” bar represents the test data and the remaining bars in between represents the training data taken from both classes at different resampling rates. For brevity bar labels are abbreviated, for example 10% SMOTE oversampling of minority class and 90% K-Medoids undersampling of majority class is labeled as “S10_K90”. A train-test dataset is formed by combining a train set and a test set (test set is kept fixed between different sampling approaches, and it need not be balanced).

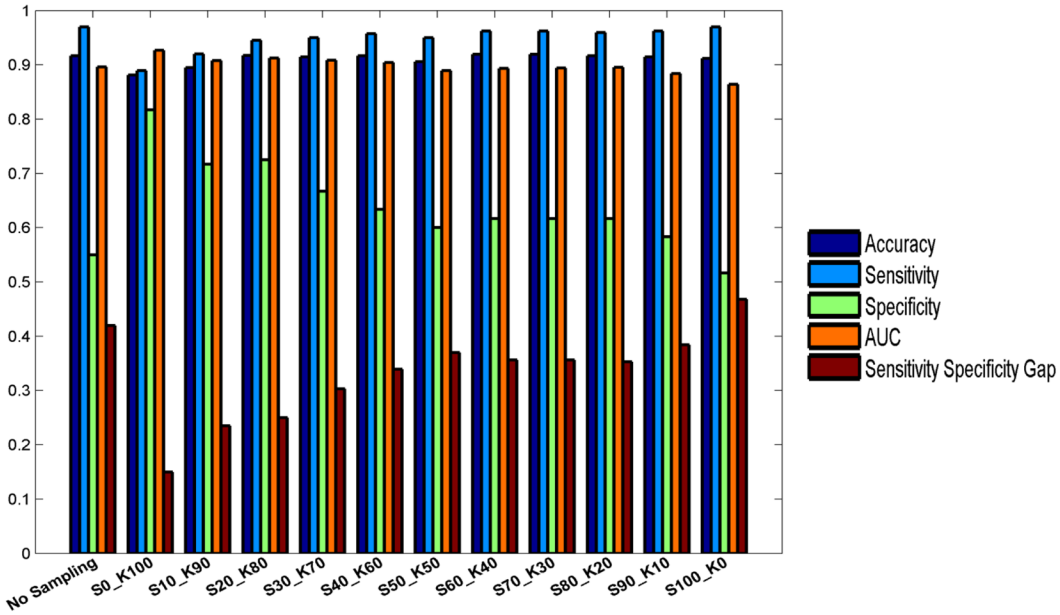

Figure 7.
NC/MCI majority voting classification performance comparison of SVM classifier, averaged across 10 cross folds, using top 10 features from SLR+SS for different rates of data sampling.

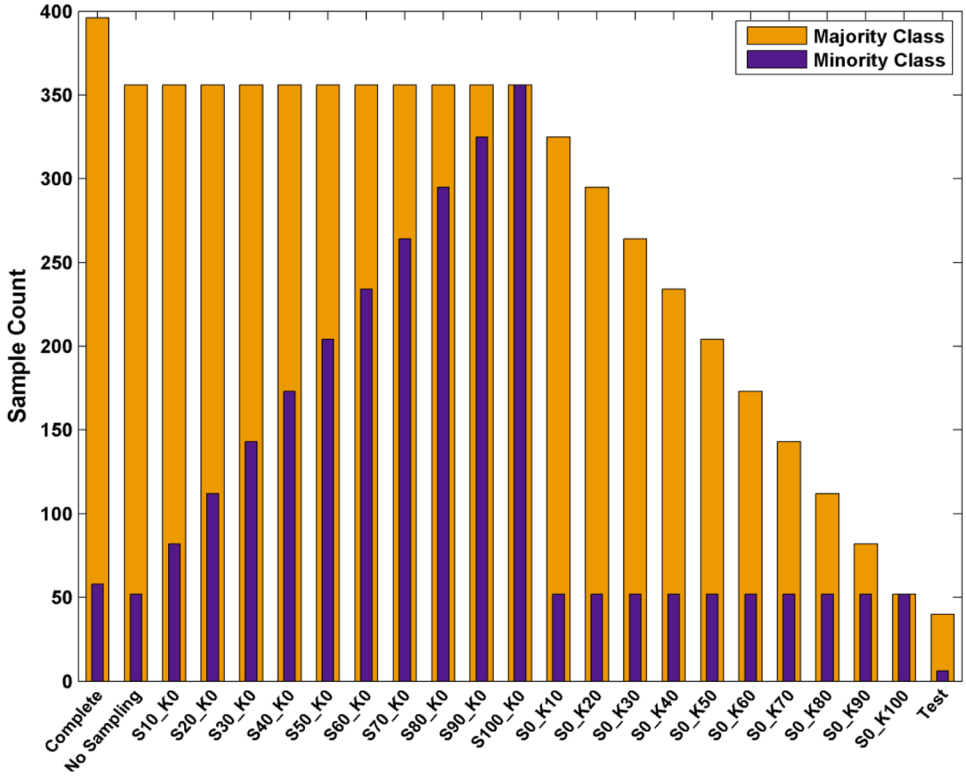

Figure 8.

The bar labeled “Complete” represents the data available for analysis. The “Test” bar represents the test data and the remaining bars in between represents the training data taken from both classes at different resampling rates. For brevity bar label are abbreviated, for example “S30_K0” corresponds to 30% SMOTE oversampling of minority class and no undersampling majority class. A train-test dataset is formed by combining a train set and a test set (test set is kept fixed between different sampling approaches, and it need not be balanced).

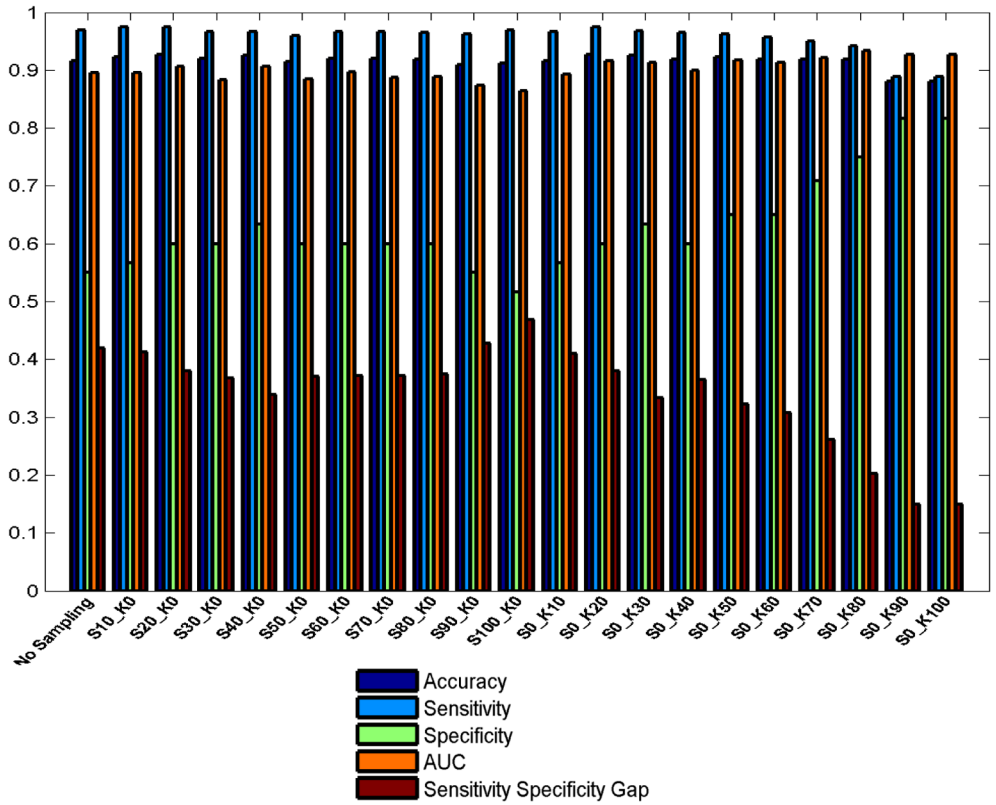

Figure 9.

NC/MCI majority voting classification performance comparison of SVM classifier, averaged across 10 cross folds, using top 10 features from SLR+SS for different rates of data sampling. Note the decreasing sensitivity-specificity gap as the rate of undersampling is increased. Complete undersampled dataset (labeled as S0_K100) showed least gap.

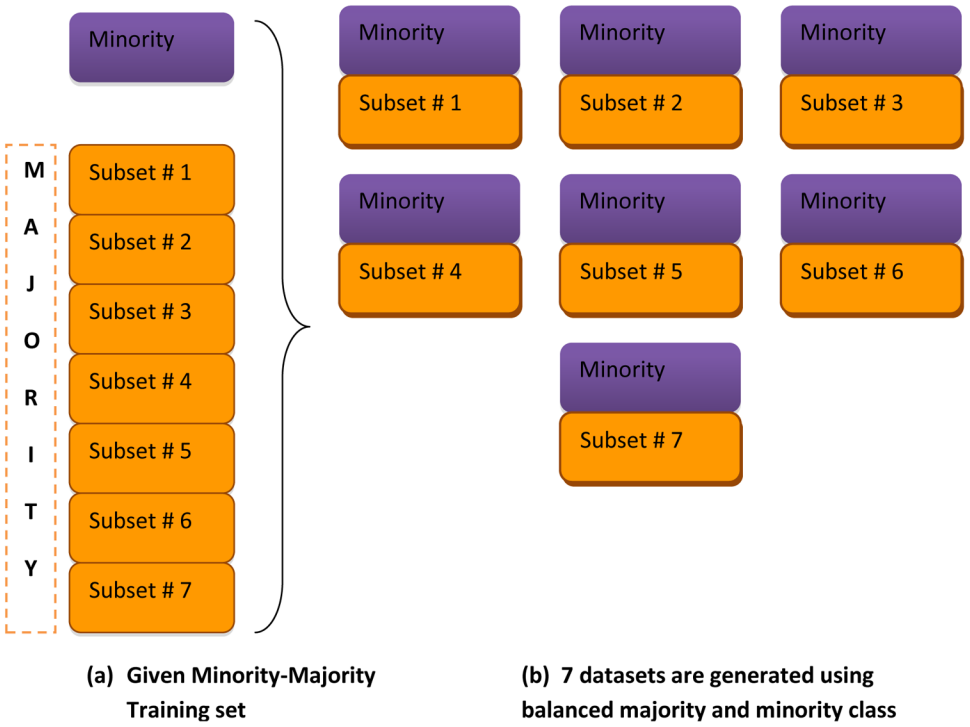

Figure 10.
Generation of classification models for imbalanced data using Chan and Stolfo (1998) approach. The majority class (represented by Orange colored rectangles in the figure) is evenly divided into minority class sized non-overlapping subsets.

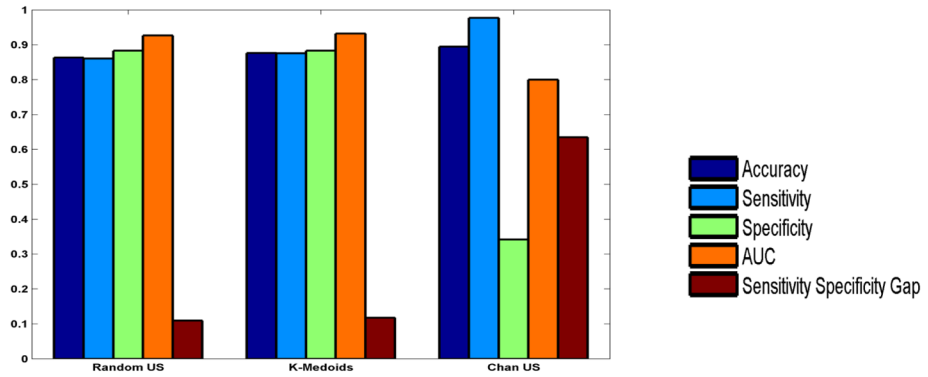

Figure 11.

NC/MCI majority voting classification performance comparison of SVM classifier for different undersampling approaches, averaged across 10 cross folds, using top 10 features from SLR+SS for different rates of data sampling depicting efficacy of K-Medoids and random undersampling approach over Chan and Stolfo proposed solution (Chan and Stolfo, 1998).

Table 1

Summary of ADNI data used in the study

ADNI Baseline Data Details		
	Proteomics	**MRI**
Feature Count	147	305
Control Cases (NC)	58	191
MCI Stable Cases	233	177
MCI Convertor Cases	163	142
AD Cases	112	138

Table 2

NC versus MCI prediction task using 147 proteomics features: Summary of data used in train-test set in each cross fold for different data re-sampling techniques. MCI includes both MCI Convertor (163) and MCI Stable (233) subjects

		No Sampling		K-Medoids/Random US		SMOTE/Random OS	
Target	**Sample #**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
NC (−)	58	52	6	52	6	351	6
MCI (+)	391	351	40	52	40	403	40
Total	**449**	**403**	**46**	**104**	**46**	**754**	**46**

Table 3

NC/MCI: Comparison of different sampling approaches using top 10 proteomics features, averaged across 10 cross folds, in terms of accuracy, sensitivity and specificity, and AUC. The best value in each column for each performance metric is underlined to compare different sampling approaches and highest value in each row is highlighted in bold to compare feature selection algorithms and classifiers

		SLR+SS				T-Test			
Accuracy (%)	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	90.152	90.152	**93.261**	**93.261**	90.620	90.620	89.717	89.717
	Random US	80.146	84.772	80.965	**86.326**	78.607	83.685	78.344	82.630
	K-Medoids	80.596	85.359	81.384	**87.630**	78.958	83.217	78.576	81.696
	Random OS	90.748	91.424	90.500	**92.293**	89.130	89.685	88.093	88.815
	SMOTE	89.971	89.902	90.761	**91.054**	87.816	88.348	88.517	89.652
Sensitivity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	**0.9850**	**0.9850**	0.9700	0.9700	**0.9850**	**0.9850**	0.9725	0.9725
	Random US	0.8017	0.8456	0.8083	**0.8608**	0.7864	0.8356	0.7818	0.8258
	K-Medoids	0.8062	0.8475	0.8127	**0.8758**	0.7910	0.8353	0.7869	0.8228
	Random OS	0.9815	**0.9897**	0.9531	0.9697	0.9663	0.9722	0.9273	0.9322
	SMOTE	0.9523	0.9522	0.9553	**0.9572**	0.9306	0.9342	0.9390	0.9492
Specificity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	0.3333	0.3333	**0.6833**	**0.6833**	0.3750	0.3750	0.3833	0.3833
	Random US	0.8033	0.8667	0.8236	**0.8833**	0.7872	0.8500	0.7983	0.8333
	K-Medoids	0.8081	**0.9000**	0.8258	0.8833	0.7828	0.8167	0.7825	0.7833
	Random OS	0.3992	0.4000	0.5717	**0.6000**	0.3775	0.3833	0.5600	0.5833
	SMOTE	0.5394	0.5333	0.5881	**0.6000**	0.5250	0.5417	0.5231	0.5417
AUC	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	0.3279	0.7997	0.6621	**0.9392**	0.3688	0.8107	0.3671	0.7924
	Random US	0.7981	0.9138	0.8114	**0.9267**	0.7816	0.9108	0.7839	0.8989
	K-Medoids	0.8007	**0.9335**	0.8129	0.9319	0.7822	0.9018	0.7808	0.8731
	Random OS	0.6629	0.7600	0.7465	**0.8317**	0.6414	0.7438	0.7253	0.8071
	SMOTE	0.7376	0.8360	0.7663	**0.8788**	0.7213	0.8313	0.7206	0.8400

Table 4

NC versus MCI prediction task using 305 MRI features: Summary of data used in train-test set in each cross fold for different data re-sampling techniques. MCI includes both MCI Convertor (142) and MCI Stable (177) subjects

		No Sampling		K-Medoids/Random US		SMOTE/Random OS	
Target	**Sample #**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
NL (−)	191	171	20	171	20	287	20
MCI (+)	319	287	32	171	32	287	32
Total	**510**	**458**	**52**	**342**	**52**	**574**	**52**

Table 5

NC/MCI: Comparison of different sampling approaches using top 10 MRI features, averaged across 10 cross folds, in terms of accuracy, sensitivity and specificity, and AUC. The best value in each column for each performance metric is underlined to compare different sampling approaches and highest value in each row is highlighted in bold to compare feature selection algorithms and classifiers

		SLR+SS				T-Test			
Accuracy (%)	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	67.436	67.436	**67.720**	**67.720**	66.044	66.044	67.482	67.482
	Random US	65.863	67.289	66.517	**69.020**	66.282	67.729	66.545	67.582
	K-Medoids	66.988	68.059	67.158	**69.496**	66.466	67.390	66.826	66.960
	Random OS	66.112	**67.866**	66.136	65.559	66.972	66.905	67.156	67.143
	SMOTE	66.001	65.128	64.871	65.321	66.357	67.051	66.808	**67.674**
Sensitivity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	0.7740	0.7740	0.7928	0.7928	0.7644	0.7644	**0.8085**	**0.8085**
	Random US	0.6312	0.6331	0.6419	**0.6518**	0.6269	0.6297	0.6129	0.6204
	K-Medoids	**0.6398**	0.6362	0.6396	0.6395	0.6203	0.6172	0.6117	0.6140
	Random OS	0.7173	**0.7178**	0.6838	0.6740	0.6996	0.6927	0.6388	0.6271
	SMOTE	**0.7004**	0.6925	0.6845	0.6893	0.6964	0.6987	0.6473	0.6549
Specificity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	**0.5136**	**0.5136**	0.4927	0.4927	0.4977	0.4977	0.4586	0.4586
	Random US	0.7134	0.7500	0.7134	0.7650	0.7323	0.7659	0.7643	**0.7800**
	K-Medoids	0.7292	0.7650	0.7343	**0.7950**	0.7495	0.7800	0.7739	0.7750
	Random OS	0.5671	0.6136	0.6273	0.6277	0.6236	0.6327	0.7278	**0.7468**
	SMOTE	0.6010	0.5918	0.5984	0.6059	0.6196	0.6359	0.7127	**0.7250**
AUC	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	0.3953	0.6984	0.3876	**0.7048**	0.3744	0.6878	0.3678	0.6873
	Random US	0.6657	0.7438	0.6708	**0.7615**	0.6729	0.7459	0.6817	0.7514
	K-Medoids	0.6769	0.7494	0.6802	**0.7715**	0.6772	0.7486	0.6856	0.7490
	Random OS	0.6184	0.6853	0.6344	0.6738	0.6391	0.6810	0.6615	**0.7103**
	SMOTE	0.6435	0.7009	0.6339	0.7028	0.6506	0.7157	0.6727	**0.7380**

Table 6

NC versus AD prediction task using 147 proteomics features. Summary of data used in train-test set in each cross fold for different data re-sampling techniques

		No Sampling		K-Medoids/Random US		SMOTE/Random OS	
Target	**Sample #**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
NL (−)	58	52	6	52	6	100	6
AD (+)	112	100	12	52	12	100	12
Total	**170**	**152**	**18**	**104**	**18**	**200**	**18**

Table 7

NC/AD: Comparison of different sampling approaches using top 10 proteomics features, averaged across 10 cross folds, in terms of accuracy, sensitivity and specificity, and AUC. The best value in each column for each performance metric is underlined to compare different sampling approaches and highest value in each row is highlighted in bold to compare feature selection algorithms and classifiers

		SLR+SS				T-Test			
Accuracy (%)	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	80.694	80.694	**84.861**	**84.861**	81.806	81.806	83.056	83.056
	Random US	78.718	83.056	80.444	**84.167**	78.995	81.250	78.500	80.833
	K-Medoids	79.037	81.806	80.690	**83.611**	77.653	80.000	78.579	80.833
	Random OS	78.861	80.278	81.639	81.389	81.000	**82.778**	80.056	81.111
	SMOTE	79.366	**81.944**	80.532	80.278	80.125	**81.944**	79.236	79.583
Sensitivity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	**0.9250**	**0.9250**	0.9167	0.9167	**0.9250**	**0.9250**	0.8917	0.8917
	Random US	0.7861	0.8167	0.8003	**0.8333**	0.7889	0.8167	0.7803	0.8083
	K-Medoids	0.7950	0.8000	0.8147	**0.8500**	0.7833	0.8000	0.7869	0.8167
	Random OS	0.8708	0.8667	0.8767	0.8750	0.8883	**0.9083**	0.8633	0.8667
	SMOTE	0.8644	**0.9000**	0.8492	0.8417	0.8772	0.8833	0.8425	0.8500
Specificity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	0.6083	0.6083	0.7250	0.7250	0.6417	0.6417	**0.7333**	**0.7333**
	Random US	0.7964	**0.8583**	0.8144	**0.8583**	0.7983	0.8167	0.7961	0.8083
	K-Medoids	0.7811	**0.8417**	0.7908	0.8083	0.7667	0.8000	0.7847	0.7917
	Random OS	0.6317	0.6750	**0.7008**	0.6917	0.6583	0.6667	0.6800	0.7000
	SMOTE	0.6700	0.6833	0.7181	**0.7250**	0.6797	0.7167	0.7042	0.7000
AUC	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	0.5569	0.8431	0.6611	**0.9125**	0.5903	0.8569	0.6528	0.8917
	Random US	0.7853	0.9056	0.8022	**0.9194**	0.7866	0.8896	0.7831	0.8847
	K-Medoids	0.7817	0.8938	0.7968	**0.9056**	0.7666	0.8715	0.7791	0.8819
	Random OS	0.7310	0.8535	0.7718	**0.9035**	0.7543	0.8778	0.7493	0.8778
	SMOTE	0.7590	0.8819	0.7782	**0.8889**	0.7733	0.8708	0.7677	0.8563

Table 8

NC versus AD prediction task using 305 MRI features. Summary of data used in train-test set in each cross fold for different data re-sampling techniques

		No Sampling		K-Medoids/Random US		SMOTE/Random OS	
Target	**Sample #**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
NL (−)	191	171	20	124	20	171	20
AD (+)	138	124	14	124	14	171	14
Total	**329**	**295**	**34**	**248**	**34**	**342**	**34**

Table 9

NC/AD: Comparison of different sampling approaches using top 10 MRI features, averaged across 10 cross folds, in terms of accuracy, sensitivity and specificity, and AUC. The best value in each column for each performance metric is underlined to compare different sampling approaches and highest value in each row is highlighted in bold to compare feature selection algorithms and classifiers

		SLR+SS				T-Test			
Accuracy (%)	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	**87.225**	**87.225**	85.908	85.908	85.460	85.460	86.343	86.343
	Random US	85.930	85.460	85.312	**87.225**	84.999	84.872	85.287	85.908
	K-Medoids	86.054	86.637	85.935	**87.379**	85.278	85.614	84.864	84.885
	Random OS	86.573	86.650	86.107	**87.097**	86.265	86.061	86.691	86.650
	SMOTE	86.306	87.225	86.140	**87.379**	85.732	86.049	85.682	86.343
Sensitivity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	**0.8262**	**0.8262**	0.8119	0.8119	0.7905	0.7905	0.7976	0.7976
	Random US	0.8413	0.8262	0.8463	**0.8476**	0.8307	0.8262	0.8306	0.8333
	K-Medoids	0.8264	0.8262	0.8377	0.8333	0.8321	**0.8405**	0.8245	0.8262
	Random OS	0.8424	**0.8536**	0.8405	0.8452	0.8329	0.8321	0.8393	0.8393
	SMOTE	0.8148	0.8262	0.8216	0.8321	0.8273	**0.8333**	0.8243	**0.8333**
Specificity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	0.9059	0.9059	0.8918	0.8918	0.9009	0.9009	**0.9109**	**0.9109**
	Random US	0.8725	0.8759	0.8583	**0.8909**	0.8632	0.8659	0.8676	0.8768
	K-Medoids	0.8851	0.8959	0.8742	**0.9018**	0.8669	0.8668	0.8638	0.8627
	Random OS	0.8855	0.8768	0.8789	**0.8918**	0.8855	0.8818	0.8884	0.8868
	SMOTE	0.8988	**0.9059**	0.8913	**0.9059**	0.8787	0.8809	0.8806	0.8859
AUC	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	**0.9849**	0.8761	0.9809	0.8616	0.9788	0.8486	0.9846	0.8564
	Random US	0.8606	0.8546	0.8562	**0.8764**	0.8511	0.8503	0.8533	0.8566
	K-Medoids	0.8606	0.8686	0.8608	**0.8798**	0.8533	0.8577	0.8490	0.8486
	Random OS	0.8752	0.8514	0.8703	0.8572	0.8717	0.8421	**0.8761**	0.8468
	SMOTE	0.8615	0.8743	0.8613	**0.8778**	0.8574	0.8625	0.8564	0.8593

Table 10

NC versus MCIC & AD prediction task using 147 proteomics features. Summary of data used in train-test set in each cross fold for different data re-sampling techniques.; MCIC & AD includes both MCI Convertor (163) and AD (112) subjects.

		No Sampling		K-Medoids/Random US		SMOTE/Random OS	
Target	**Sample #**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
NL (−)	58	52	6	52	6	247	6
MCIC & AD (+)	275	247	28	52	28	247	28
Total	**333**	**299**	**34**	**104**	**34**	**494**	**34**

Table 11

NC/MCIC & AD: Comparison of different sampling approaches using top 10 proteomics features, averaged across 10 cross folds, in terms of accuracy, sensitivity and specificity, and AUC. The best value in each column for each performance metric is underlined to compare different sampling approaches and highest value in each row is highlighted in bold to compare feature selection algorithms and classifiers.

		SLR+SS				T-Test			
Accuracy (%)	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	88.224	88.224	**89.771**	**89.771**	88.301	88.301	87.865	87.865
	Random US	78.350	**83.965**	78.782	83.889	77.395	80.795	78.112	79.837
	K-Medoids	78.499	83.671	79.038	**84.771**	77.648	81.166	78.474	83.224
	Random OS	84.771	86.536	85.436	86.024	84.962	86.242	85.565	**87.418**
	SMOTE	85.629	86.536	86.822	**86.830**	86.355	86.318	86.118	86.536
Sensitivity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	**0.9857**	**0.9857**	0.9707	0.9707	**0.9857**	**0.9857**	0.9679	0.9679
	Random US	0.7822	0.8270	0.7870	**0.8370**	0.7684	0.8033	0.7774	0.7953
	K-Medoids	0.7834	0.8306	0.7924	**0.8512**	0.7767	0.8076	0.7870	0.8326
	Random OS	0.9162	0.9227	0.9000	0.9020	0.9255	**0.9370**	0.8841	0.8941
	SMOTE	0.9234	0.9314	**0.9316**	0.9314	0.9232	0.9242	0.9195	0.9234
Specificity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	0.3833	0.3833	**0.5500**	**0.5500**	0.3917	0.3917	0.4583	0.4583
	Random US	0.7906	**0.9000**	0.7922	0.8500	0.8022	0.8333	0.8011	0.8167
	K-Medoids	0.7931	**0.8667**	0.7803	0.8333	0.7781	0.8333	0.7767	0.8333
	Random OS	0.5300	0.6000	0.6425	0.6667	0.4975	0.5167	0.7250	**0.7833**
	SMOTE	0.5361	0.5500	0.5656	0.5667	0.5828	0.5750	0.5881	**0.5917**
AUC	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	0.3774	0.8121	0.5300	**0.8823**	0.3857	0.8148	0.4399	0.8310
	Random US	0.7800	**0.9149**	0.7827	0.9068	0.7804	0.8776	0.7814	0.8851
	K-Medoids	0.7808	0.9099	0.7781	**0.9208**	0.7705	0.8808	0.7779	0.9091
	Random OS	0.6977	0.8357	0.7512	0.8428	0.6818	0.7941	0.7863	**0.9030**
	SMOTE	0.7213	**0.8561**	0.7438	0.8466	0.7448	0.8521	0.7471	0.8477

Table 12

NC versus MCIC & AD prediction task using 305 MRI features. Summary of data used in train-test set in each cross fold for different data re-sampling techniques.; MCIC & AD includes both MCI Convertor (142) and AD (138) subjects

		No Sampling		K-Medoids/Random US		SMOTE/Random OS	
Target	**Sample #**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
NL (−)	191	171	20	171	20	252	20
MCIC & AD (+)	280	252	28	171	28	252	28
Total	**471**	**423**	**48**	**342**	**48**	**504**	**48**

Table 13

NC/MCIC & AD: Comparison of different sampling approaches using top 10 MRI features, averaged across 10 cross folds, in terms of accuracy, sensitivity and specificity, and AUC. The best value in each column for each performance metric is underlined to compare different sampling approaches and highest value in each row is highlighted in bold to compare feature selection algorithms and classifiers

		SLR+SS				T-Test			
Accuracy (%)	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	85.321	85.321	**85.529**	**85.529**	82.356	82.356	83.446	83.446
	Random US	83.888	84.904	84.424	**85.112**	82.540	83.237	82.177	81.939
	K-Medoids	83.735	84.279	84.216	**85.112**	82.603	83.446	82.668	83.029
	Random OS	84.014	83.718	83.681	**85.272**	83.268	83.141	82.901	82.516
	SMOTE	85.091	85.529	85.193	**86.362**	82.913	82.612	82.451	82.612
Sensitivity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	0.8750	0.8750	**0.8786**	**0.8786**	0.8429	0.8429	0.8607	0.8607
	Random US	0.8207	**0.8286**	0.8273	0.8179	0.8062	0.8107	0.8027	0.8036
	K-Medoids	0.8212	0.8250	**0.8267**	0.8250	0.8113	0.8214	0.8121	0.8179
	Random OS	0.8218	0.8179	0.8225	**0.8250**	0.8061	0.8071	0.7932	0.7857
	SMOTE	0.8537	0.8571	0.8619	**0.8750**	0.8246	0.8214	0.8190	0.8214
Specificity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	**0.8250**	**0.8250**	**0.8250**	**0.8250**	0.7959	0.7959	0.8000	0.8000
	Random US	0.8666	0.8800	0.8709	**0.9000**	0.8554	0.8650	0.8515	0.8450
	K-Medoids	0.8625	0.8700	0.8666	**0.8900**	0.8497	0.8550	0.8498	0.8500
	Random OS	0.8645	0.8618	0.8519	**0.8909**	0.8697	0.8659	0.8794	0.8809
	SMOTE	0.8493	**0.8550**	0.8390	0.8500	0.8378	0.8350	0.8345	0.8350
AUC	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	None	0.7236	0.8612	0.7273	**0.8658**	0.6737	0.8333	0.6930	0.8437
	Random US	0.8397	0.8655	0.8452	**0.8757**	0.8265	0.8530	0.8222	0.8393
	K-Medoids	0.8374	0.8612	0.8422	**0.8728**	0.8254	0.8505	0.8261	0.8450
	Random OS	0.8303	0.8736	0.8233	**0.8869**	0.8246	0.8770	0.8222	0.8699
	SMOTE	0.8481	0.8660	0.8464	**0.8805**	0.8266	0.8425	0.8222	0.8407

Table 14

NC versus MCI prediction task using 147 proteomics features. Summary of data used in train-test set in each cross fold for combination resampling techniques. MCI includes both MCI Convertor (163) and MCI Stable (233) subjects

SMOTE %		**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**	
K-Medoids %		**100%**	**90%**	**80%**	**70%**	**60%**	**50%**	**40%**	**30%**	**20%**	**10%**	**0%**	
Target	**#**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Test**
NL(−)	58	52	82	112	143	173	204	234	264	295	325	356	6
MCI(+)	396	52	82	112	143	173	204	234	264	295	325	356	40
Total	**454**	**104**	**164**	**224**	**286**	**346**	**408**	**468**	**528**	**590**	**650**	**712**	**46**

Table 15

NC/MCI: Comparison of different sampling approaches using top 10 proteomics features obtained by SLR+SS and T-Test, averaged across 10 cross folds, in terms of accuracy, sensitivity and specificity, and AUC. The best value in each column for each performance metric is underlined to compare different sampling approaches and highest value in each row is highlighted in bold to compare feature selection algorithms and classifiers. OS% refers to SMOTE oversampling percentage and US% corresponds to K-Medoids undersampling percentage. Results obtained without resampling approach are indicated by row (0%,0%), (0%,100%) refers to complete undersampling, and (100%,0%) corresponds to complete oversampling

		SLR+SS				T-Test			
Accuracy (%)	**(OS%,US%)**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	(0%,0%)	94.022	**94.565**	92.283	92.391	89.130	89.130	90.217	90.217
	(0%,100%)	80.596	85.359	81.384	**87.630**	78.958	83.217	78.576	81.696
	(10%,90%)	83.587	**90.217**	85.217	89.130	83.261	85.870	84.674	**90.217**
	(20%,80%)	84.348	88.043	87.283	89.130	86.304	**90.217**	85.761	**90.217**
	(30%,70%)	87.609	90.217	88.804	**91.304**	87.500	89.130	88.370	**91.304**
	(40%,60%)	87.391	89.130	90.435	**92.391**	86.522	89.130	86.630	89.130
	(50%,50%)	88.261	89.130	89.022	**90.217**	88.804	89.130	89.565	**90.217**
	(60%,40%)	88.478	89.130	89.565	**91.304**	88.261	90.217	87.935	88.043
	(70%,30%)	87.717	89.130	89.674	**92.391**	88.043	89.130	87.935	88.043
	(80%,20%)	87.935	88.043	90.109	**92.391**	89.457	90.217	89.457	89.130
	(90%,10%)	87.935	88.043	91.087	**92.391**	89.022	90.217	88.804	89.130
	(100%,0%)	89.971	89.902	90.761	**91.054**	87.816	88.348	88.517	89.652
Sensitivity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	(0%,0%)	0.984	**0.988**	0.948	0.950	**0.988**	**0.988**	0.975	0.975
	(0%,100%)	0.806	0.848	0.813	**0.876**	0.791	0.835	0.787	0.823
	(10%,90%)	0.833	0.900	0.855	0.900	0.835	0.863	0.844	**0.913**
	(20%,80%)	0.854	0.888	0.885	0.900	0.879	**0.925**	0.873	0.913
	(30%,70%)	0.891	**0.925**	0.905	**0.925**	0.893	0.913	0.900	**0.925**
	(40%,60%)	0.901	0.925	0.924	**0.938**	0.885	0.913	0.890	0.913
	(50%,50%)	0.913	**0.925**	0.915	**0.925**	0.908	0.913	0.914	**0.925**
	(60%,40%)	0.918	0.925	0.923	**0.938**	0.910	0.925	0.906	0.913
	(70%,30%)	0.909	0.925	0.924	**0.950**	0.904	0.913	0.905	0.913
	(80%,20%)	0.913	0.913	0.928	**0.950**	0.916	0.925	0.923	0.925

		SLR+SS				T-Test			
	(90%,10%)	0.911	0.913	0.939	**0.950**	0.914	0.925	0.919	0.913
	(100%,0%)	0.952	0.952	0.955	**0.957**	0.931	0.934	0.939	0.949
Specificity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	(0%,0%)	0.650	0.667	**0.758**	0.750	0.250	0.250	0.417	0.417
	(0%,100%)	0.808	**0.900**	0.826	0.883	0.783	0.817	0.783	0.783
	(10%,90%)	0.858	**0.917**	0.833	0.833	0.817	0.833	0.867	0.833
	(20%,80%)	0.775	**0.833**	0.792	**0.833**	0.758	0.750	0.758	**0.833**
	(30%,70%)	0.775	0.750	0.775	**0.833**	0.758	0.750	0.775	**0.833**
	(40%,60%)	0.692	0.667	0.775	**0.833**	0.733	0.750	0.708	0.750
	(50%,50%)	0.683	0.667	0.725	0.750	0.758	0.750	**0.775**	0.750
	(60%,40%)	0.667	0.667	0.717	**0.750**	0.700	**0.750**	0.700	0.667
	(70%,30%)	0.667	0.667	0.717	**0.750**	0.725	**0.750**	0.708	0.667
	(80%,20%)	0.658	0.667	0.725	**0.750**	**0.750**	**0.750**	0.708	0.667
	(90%,10%)	0.667	0.667	0.725	**0.750**	0.733	**0.750**	0.683	**0.750**
	(100%,0%)	0.539	0.533	0.588	**0.600**	0.525	0.542	0.523	0.542
AUC	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	(0%,0%)	0.801	0.869	0.841	**0.900**	0.248	0.654	0.413	0.692
	(0%,100%)	0.801	**0.933**	0.813	0.932	0.782	0.902	0.781	0.873
	(10%,90%)	0.825	**0.946**	0.829	0.933	0.796	0.883	0.842	0.896
	(20%,80%)	0.800	0.908	0.830	**0.915**	0.803	0.848	0.794	0.896
	(30%,70%)	0.811	0.848	0.818	**0.927**	0.807	0.844	0.817	0.898
	(40%,60%)	0.781	0.833	0.832	**0.938**	0.799	0.844	0.789	0.846
	(50%,50%)	0.782	0.833	0.801	**0.885**	0.811	0.844	0.825	0.848
	(60%,40%)	0.776	0.833	0.805	**0.883**	0.784	0.848	0.791	0.838
	(70%,30%)	0.773	0.833	0.796	**0.894**	0.800	0.842	0.790	0.838
	(80%,20%)	0.766	0.825	0.817	**0.894**	0.819	0.848	0.807	0.848
	(90%,10%)	0.768	0.825	0.821	**0.894**	0.811	0.848	0.777	0.846
	(100%,0%)	0.738	0.836	0.766	**0.879**	0.721	0.831	0.721	0.840

Table 16

NC versus MCI prediction task using 147 proteomics features. Summary of data used in train-test set in each cross fold for different rates of oversampling. MCI includes both MCI Convertor (163) and MCI Stable (233) subjects

SMOTE %		**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**	
Target	**Count**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Test**
NL(−)	58	52	82	112	143	173	204	234	264	295	325	356	6
MCI(+)	396	356	356	356	356	356	356	356	356	356	356	356	40
Total	**454**	**408**	**438**	**468**	**499**	**529**	**560**	**590**	**620**	**651**	**681**	**712**	**46**

Table 17

NC versus MCI prediction task using 147 proteomics features. Summary of data used in train-test set in each cross fold for different rates of undersampling. MCI includes both MCI Convertor (163) and MCI Stable (233) subjects

K-Medoids %		**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**	
Target	**Count**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Test**
NL(−)	58	52	52	52	52	52	52	52	52	52	52	52	6
MCI(+)	396	356	325	295	264	234	204	173	143	112	82	52	40
Total	**454**	**408**	**377**	**347**	**316**	**286**	**256**	**225**	**195**	**164**	**134**	**104**	**46**

Table 18

NC/MCI: Comparison of different sampling approaches using top 10 proteomics features obtained by SLR+SS and T-Test, averaged across 10 cross folds, in terms of accuracy, sensitivity and specificity, and AUC. The best value in each column for each performance metric is underlined to compare different sampling approaches and highest value in each row is highlighted in bold to compare feature selection algorithms and classifiers. OS% refers to SMOTE oversampling percentage and US% corresponds to K-Medoids undersampling percentage. Results obtained without resampling approach are indicated by row (0%,0%), (0%,100%) refers to complete undersampling, and (100%,0%) corresponds to complete oversampling

		SLR+SS				T-Test			
Accuracy (%)	**(OS%,US%)**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	(0%,0%)	94.022	**94.565**	92.283	92.391	90.000	89.130	90.000	90.217
	(10%,0%)	91.848	92.391	93.804	**94.565**	88.370	88.043	90.543	89.130
	(20%,0%)	90.217	89.130	92.500	**93.478**	89.348	90.217	91.304	90.217
	(30%,0%)	90.543	90.217	91.630	**92.391**	89.891	89.130	89.783	89.130
	(40%,0%)	88.913	90.217	91.522	**92.391**	89.565	90.217	88.913	89.130
	(50%,0%)	89.130	89.130	**91.630**	91.304	89.348	89.130	89.022	89.130
	(60%,0%)	89.457	90.217	91.957	**92.391**	89.565	89.130	89.130	88.043
	(70%,0%)	89.239	89.130	90.652	**92.391**	89.457	89.130	89.565	89.130
	(80%,0%)	87.717	88.043	90.326	**91.304**	90.000	90.217	89.022	89.130
	(90%,0%)	87.609	88.043	89.348	**90.217**	90.000	**90.217**	89.674	89.130
	(100%,0%)	90.036	88.043	89.638	90.217	91.993	**93.478**	90.870	91.304
	(0%,10%)	93.043	**93.478**	93.043	**93.478**	90.217	90.217	89.348	90.217
	(0%,20%)	93.370	93.478	93.152	**94.565**	89.783	90.217	89.457	89.130
	(0%,30%)	92.717	**93.478**	92.283	92.391	89.457	90.217	88.913	90.217
	(0%,40%)	**92.500**	92.391	91.630	91.304	89.565	89.130	88.478	88.043
	(0%,50%)	93.261	**93.478**	92.283	**93.478**	89.022	89.130	88.804	88.043
	(0%,60%)	92.500	**93.478**	91.196	91.304	89.674	89.130	90.652	90.217
	(0%,70%)	92.174	**93.478**	89.022	91.304	89.130	90.217	88.587	90.217
	(0%,80%)	89.457	90.217	89.348	**94.565**	88.261	89.130	88.152	90.217
	(0%,90%)	80.000	84.783	82.065	88.043	85.326	89.130	85.870	**90.217**
	(0%,100%)	81.522	84.783	82.355	**90.217**	79.420	83.696	79.493	82.609
Sensitivity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	(0%,0%)	0.984	0.988	0.948	0.950	**0.989**	0.988	0.975	0.975

		SLR+SS				T-Test			
	(10%,0%)	0.968	**0.975**	0.959	0.963	0.958	0.950	0.960	0.950
	(20%,0%)	0.949	0.938	0.951	**0.963**	0.948	0.950	0.955	0.938
	(30%,0%)	0.944	0.938	0.948	**0.963**	0.944	0.938	0.938	0.925
	(40%,0%)	0.928	0.938	0.944	**0.950**	0.936	0.938	0.930	0.925
	(50%,0%)	0.928	0.925	**0.941**	0.938	0.928	0.925	0.926	0.925
	(60%,0%)	0.929	0.938	0.949	**0.950**	0.930	0.925	0.933	0.925
	(70%,0%)	0.926	0.925	0.934	**0.950**	0.925	0.925	0.925	0.925
	(80%,0%)	0.910	0.913	0.928	**0.938**	0.926	0.925	0.923	0.925
	(90%,0%)	0.908	0.913	0.916	0.925	0.923	0.925	**0.926**	0.913
	(100%,0%)	0.928	0.913	0.920	0.925	0.942	**0.963**	0.939	0.950
	(0%,10%)	0.983	**0.988**	0.953	0.950	0.986	**0.988**	0.970	0.975
	(0%,20%)	0.986	**0.988**	0.954	0.963	0.976	0.975	0.966	0.963
	(0%,30%)	**0.988**	**0.988**	0.948	0.950	0.974	0.975	0.961	0.963
	(0%,40%)	**0.980**	0.975	0.940	0.938	0.965	0.963	0.954	0.950
	(0%,50%)	0.983	**0.988**	0.944	0.950	0.958	0.963	0.948	0.950
	(0%,60%)	0.970	**0.975**	0.936	0.938	0.955	0.950	0.953	0.950
	(0%,70%)	0.961	**0.975**	0.908	0.925	0.939	0.950	0.926	0.938
	(0%,80%)	0.920	0.925	0.903	**0.950**	0.913	0.925	0.896	0.913
	(0%,90%)	0.788	0.838	0.811	0.875	0.873	**0.913**	0.873	**0.913**
	(0%,100%)	0.797	0.825	0.804	**0.888**	0.775	0.813	0.775	0.800
Specificity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	(0%,0%)	0.650	0.667	**0.758**	0.750	0.308	0.250	0.400	0.417
	(10%,0%)	0.592	0.583	0.800	**0.833**	0.392	0.417	0.542	0.500
	(20%,0%)	0.592	0.583	**0.750**	**0.750**	0.533	0.583	0.633	0.667
	(30%,0%)	0.650	0.667	**0.708**	0.667	0.600	0.583	0.633	0.667
	(40%,0%)	0.633	0.667	0.725	**0.750**	0.625	0.667	0.617	0.667
	(50%,0%)	0.650	0.667	**0.750**	**0.750**	0.667	0.667	0.650	0.667
	(60%,0%)	0.667	0.667	0.725	**0.750**	0.667	0.667	0.617	0.583
	(70%,0%)	0.667	0.667	0.725	**0.750**	0.692	0.667	0.700	0.667

		SLR+SS				T-Test			
	(80%,0%)	0.658	0.667	0.742	**0.750**	0.725	**0.750**	0.675	0.667
	(90%,0%)	0.667	0.667	0.742	**0.750**	**0.750**	**0.750**	0.700	**0.750**
	(100%,0%)	0.719	0.667	0.739	0.750	**0.772**	0.750	0.708	0.667
	(0%,10%)	0.583	0.583	0.783	**0.833**	0.342	0.333	0.383	0.417
	(0%,20%)	0.583	0.583	0.783	**0.833**	0.375	0.417	0.417	0.417
	(0%,30%)	0.525	0.583	**0.758**	0.750	0.367	0.417	0.408	0.500
	(0%,40%)	0.558	0.583	**0.758**	0.750	0.433	0.417	0.425	0.417
	(0%,50%)	0.600	0.583	0.783	**0.833**	0.442	0.417	0.492	0.417
	(0%,60%)	0.625	0.667	**0.750**	**0.750**	0.508	0.500	0.600	0.583
	(0%,70%)	0.658	0.667	0.775	**0.833**	0.575	0.583	0.617	0.667
	(0%,80%)	0.725	0.750	0.833	**0.917**	0.683	0.667	0.783	0.833
	(0%,90%)	0.883	**0.917**	0.883	**0.917**	0.725	0.750	0.767	0.833
	(0%,100%)	0.936	**1.000**	0.956	**1.000**	0.925	**1.000**	0.928	**1.000**
AUC	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	(0%,0%)	0.801	0.869	0.841	**0.900**	0.615	0.654	0.652	0.692
	(10%,0%)	0.754	0.813	0.861	**0.956**	0.638	0.683	0.731	0.731
	(20%,0%)	0.751	0.792	0.831	**0.904**	0.717	0.800	0.784	0.833
	(30%,0%)	0.776	0.844	0.812	**0.856**	0.749	0.792	0.765	0.840
	(40%,0%)	0.761	0.844	0.822	**0.894**	0.755	0.840	0.759	0.840
	(50%,0%)	0.767	0.833	0.835	**0.894**	0.777	0.833	0.772	0.840
	(60%,0%)	0.779	0.844	0.823	**0.894**	0.780	0.833	0.748	0.833
	(70%,0%)	0.778	0.833	0.815	**0.894**	0.791	0.833	0.804	0.840
	(80%,0%)	0.762	0.825	0.823	**0.885**	0.816	0.848	0.787	0.840
	(90%,0%)	0.768	0.825	0.808	**0.875**	0.822	0.848	0.790	0.846
	(100%,0%)	0.818	0.873	0.823	**0.954**	0.852	0.892	0.818	0.879
	(0%,10%)	0.762	0.817	0.861	**0.946**	0.635	0.654	0.640	0.704
	(0%,20%)	0.767	0.817	0.856	**0.956**	0.643	0.692	0.653	0.690
	(0%,30%)	0.724	0.817	0.829	**0.900**	0.640	0.692	0.647	0.738
	(0%,40%)	0.740	0.813	0.830	**0.898**	0.670	0.685	0.658	0.683

		SLR+SS				T-Test			
	(0%,50%)	0.777	0.817	0.844	**0.956**	0.668	0.685	0.677	0.683
	(0%,60%)	0.775	0.865	0.833	**0.898**	0.704	0.742	0.750	0.790
	(0%,70%)	0.791	0.865	0.826	**0.935**	0.739	0.800	0.745	0.840
	(0%,80%)	0.803	0.892	0.857	**0.969**	0.766	0.833	0.821	0.896
	(0%,90%)	0.821	0.894	0.832	**0.935**	0.782	0.844	0.801	0.896
	(0%,100%)	0.864	0.967	0.874	0.973	0.850	**0.977**	0.850	**0.977**

Table 19

NC/MCI: Comparison of undersampling approach using top 10 proteomics features obtained by SLR+SS and T-Test, averaged across 10 cross folds, in terms of accuracy, sensitivity and specificity, and AUC. The best value in each column for each performance metric is underlined to compare different sampling approaches and highest value in each row is highlighted in bold to compare feature selection algorithms and classifiers. Chan US refers to undersampling using Chan et al.(Chan and Stolfo, 1998) approach

		SLR+SS				T-Test			
Accuracy (%)	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	Random US	80.146	84.772	80.965	**86.326**	78.607	83.685	78.344	82.630
	K-Medoids	80.596	85.359	81.384	**87.630**	78.958	83.217	78.576	81.696
	Chan US	87.210	**91.304**	87.030	89.467	86.284	89.250	85.787	90.087
Sensitivity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	Random US	0.802	0.846	0.808	**0.861**	0.786	0.836	0.782	0.826
	K-Medoids	0.806	0.848	0.813	**0.876**	0.791	0.835	0.787	0.823
	Chan US	0.942	**0.985**	0.937	0.977	0.932	0.972	0.923	0.964
Specificity	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	Random US	0.803	0.867	0.824	**0.883**	0.787	0.850	0.798	0.833
	K-Medoids	0.808	**0.900**	0.826	0.883	0.783	0.817	0.783	0.783
	Chan US	0.398	0.433	0.419	0.342	0.393	0.350	0.420	**0.475**
AUC	**Sampling Type**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**	**RF Avg**	**RF MajVote**	**SVM Avg**	**SVM MajVote**
	Random US	0.798	0.914	0.811	**0.927**	0.782	0.911	0.784	0.899
	K-Medoids	0.801	**0.933**	0.813	0.932	0.782	0.902	0.781	0.873
	Chan US	0.619	**0.830**	0.620	0.800	0.611	0.771	0.626	0.808