MACHINE LEARNING APPROACHES FOR THE COMPUTER AIDED

DIAGNOSIS AND PREDICTION OF ALZHEIMER'S DISEASE BASED ON

CLINICAL DATA

BY

RAHILA UMER

(Under the direction of Khaled Rasheed)

ABSTRACT

Alzheimer's is an irreversible brain disease that impairs memory, thinking, and behavior
and leads, ultimately, to death. It is a major public health problem in the elder population
and has a huge impact on society. It is useful to diagnose AD as early as possible, in
order to improve the quality of life of the patient and their care takers. In this thesis we
analyze the performance of different machine learning methods for the task of classifying
different subject's cognitive status as normal (NL), mild cognitive impairment (MCI) or
Alzheimer's disease (AD). Machine learning methods are used to improve the
performance of computer aided diagnosis of Alzheimer's disease using clinical data.
Also, feature subset selection methods are used to eliminate redundant attributes and
retain the most relevant features. The results are very promising and demonstrate the
utility of machine learning methods in this domain. The second part of the study is to
predict the possible conversion from MCI to AD. We conducted many experiments with

various learning algorithms and achieved performance levels comparable to the previously published results.

INDEX KEYWORDS: Machine learning Algorithms, Alzheimer's disease, Mild cognitive Impairment, Support vector machine, neural network, Feature subset selection, Meta cost learner, AD

MACHINE LEARNING APPROACHES FOR THE COMPUTER AIDED

DIAGNOSIS AND PREDICTION OF ALZHEIMER'S DISEASE BASED ON

CLINICAL DATA

BY

RAHILA UMER

B.S. Balochistan University of Information Technology and Engineering Management

Sciences, BUITEMS, Pakistan, 2006

A Thesis Submitted to the Graduate Faculty of the University of Georgia in Partial
Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2011

MACHINE LEARNING APPROACHES FOR THE COMPUTER AIDED

DIAGNOSIS AND PREDICTION OF ALZHEIMER'S DISEASE BASED ON

CLINICAL DATA

By

RAHILA UMER

Major Professor: Khaled Rasheed

Committee: Tianming Liu

                     Hamid R. Arabnia

Electronic Version Approved:

Maureen Grasso

Dean of the Graduate School

The University of Georgia

August 2011

DEDICATION

To my loving Parents and my Brothers

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION AND LITERATURE REVIEW

Alzheimer's disease (AD) is the most common form of dementia. This incurable, degenerative and terminal disease was first described by German psychiatrist and Neuropathologist Alois Alzheimer in 1906 and was named after him [1]. The risk of developing AD increases with age. It is most often diagnosed in people over 65 years of age. Nevertheless, AD is not always a disease of old age; many individuals younger than age 65 can also develop the disease. It is irreversible, and progressively destroys memory and thinking skills which results in decline of memory and mental function. Symptoms include confusion, irritability, aggression, mood swings, language breakdown, long-term memory loss and the decline of sufferer's senses. Ultimately this leads to the loss of bodily functions and death [2].

Scientists estimate that up to 4 million people now have AD in the USA alone and for every 5-year age group beyond 65 the percentage doubles. According to the Alzheimer's Association there are 500,000 Americans younger than 65 with Alzheimer's and other dementias, of which 40 percent are estimated to have Alzheimer's [2]. If no preventive measures or treatments are taken, it is estimated that 14 million older Americans will have Alzheimer's disease in 50 years [3]. AD and other forms of dementia have huge impact on the US economy, costing more than $100 billion per year.

Current treatments cannot stop Alzheimer's from progressing but can slow down the worsening of symptoms. Early diagnosis of AD can help improve the quality of life of the patients and their families; it also helps researchers to deeply understand the causes of the disease to reverse or slow down the progress of AD and offers more chances to treatments in the early stages. The clinical diagnosis of Alzheimer's is based on the investigation of the complete medical history, conducting lab tests, physical exam and neuropsychological tests that measure memory, attention, language skills and problem-solving abilities. Accurate diagnosis of Alzheimer's disease is challenging since there are other causes of dementia that could have the same symptoms. Severe cognitive deficit and autopsy confirmation of histopathological changes in the brain confirms the diagnosis of AD. For a living person the diagnosis confirms if the deficits are severe enough that they interfere with normal daily functions. Recent studies show that AD has a pre-symptomatic phase likely lasting for years, and during this phase there is a high probability of preserving the cognitive functions through proper treatment [4]. However during this stage clinical symptoms are not apparent. These early signs of the degenerative process that are most likely to evolve to AD are characterized as mild cognitive impairment (MCI). Subjects with MCI have high risk of AD and it is considered as a transitional zone between normal aging and AD [5].

Machine learning and computer-aided diagnosis (CAD) have gained increasing attention in the medical field. A machine learning algorithm is trained using a set of examples to produce a desired output. Those examples are divided into different classes. When a new instance is presented to the learning algorithm, it assigns it a class according to the set of classification rules. The only information given to the algorithm is a set of labeled

examples (i.e. a set of instances with their class). These classification algorithms induce the classification rules from the data.

The objective of this study is to improve the diagnostic performance of current diagnostic methods for AD by evaluating all clinical diagnostic information with supervised machine learning algorithms. We study the performance of machine learning methods on a clinical dataset that is extracted from the Alzheimer's disease Neuroimaging Initiative (ADNI) database.

The first part of the thesis concentrates on the analysis of machine learning methods on the task of classifying different subject's cognitive status as NL (normal), MCI (mild cognitive impairment) or AD (Alzheimer disease). The following tasks are performed in first part:

1. The performance characteristics of different machine learning algorithms are compared based on classification accuracy.
2. Feature subset selection methods are used to eliminate the redundant attributes.
3. The performance characteristics of classifiers are compared with and without feature selection methods.

The focus of the second part of the thesis is on predicting the possible conversion from MCI to probable AD. Following subtasks are performed in second part

1. The Performance characteristics of different machine learning prediction algorithms are compared based on sensitivity and specificity.
2. To manage the tradeoff between specificity and sensitivity, Meta cost classifiers are used and their performance is analyzed.

1.1 PREVIOUS WORK

This section briefly surveys the previous work done in this domain, i.e., the use of machine learning methods for the improvement of the diagnosis of AD.

The Classification of different stages of Alzheimer's disease using machine learning methods was addressed in [7]. Stages of AD are divided as Mild, Moderate and Severe AD. The dataset used in that paper was collected at the National Institute on Aging NIA. Many machine learning algorithms are compared based on accuracy and run time. For the evaluation of the classifier they used a testing set. In that study, the highest accuracy achieved was 99.55%.

In [8] a predictive model is developed using machine learning methods where focus is to identify the possible conversion from MCI to AD. The dataset used in the study was from the ADNI database. They used the Area under the ROC curve (AUC) as a metric of performance measurement for the classification. The highest accuracy they got was 0.887 AUC. However they are using an oversampling technique called SMOTE which produces new (hypothetical) instances for the rare class and which may not be authentic in a biomedical domain.

In [9] Support vector machines (SVM) are used to distinguish between subjects as AD or elderly control subjects by using whole brain Magnetic Resonance Images (MRI). The dataset was comprised of 16 subjects with AD and 22 elderly control subjects. The

highest classification accuracy obtained was 94.5% with specificity of 96.6% and sensitivity of 91.5%.

In [10] the authors proposed a computer aided diagnosis system for the early diagnosis of Alzheimer's disease using Single Emission Computed Tomography (SPECT) images. The proposed method is based on random forests as a predictor. With the help of feature extraction algorithms, the highest accuracy achieved is 96.9% with sensitivity of 100% and specificity of 92.7%.

In [22] Support vector machines with an SVM based feature selection method were trained to differentiate between AD and healthy controls, then this trained model is used to predict possible conversion from MCI to AD. This classification was based on the Structural Magnetic Resonance Imaging data. The highest accuracy achieved was 90.5% for classifying AD and healthy control, and 72.3% accuracy for predicting MCI conversion to AD.

Chapter 2

BACKGROUND INFORMATION

2.1   ADNI CLINICAL DATASET

The data used in this study is obtained from the Alzheimer's disease Neuroimaging
Initiative (ADNI) database. ADNI was launched by the National Institute on Aging
(NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and
the Food and Drug Administration (FDA) in 2003. The objective of this project is to
study the rate of change of cognition, function, and brain structure to define the rate of
progressive MCI and AD. ADNI uses neuroimaging and biomarker measures to track the
changes in the brains of the study subjects to diagnose AD at an initial stage. ADNI
recruits people who are healthy but with a memory problem or diagnosed with late-stage
MCI or with early Alzheimer's disease and the participant ages range from 55 to 90.
ADNI recruited 800 subjects where 200 were cognitively healthy, 400 with MCI and 200
with early AD. The subjects are being followed for about two to three years. AD subjects
will be studied at 0, 6, 12 and 24 months. MCI subjects that are at high risk of conversion
to AD are studied at 0, 6, 12, 18, 24 and 36 months. ADNI collects the data of the
subjects during the study and maintains a database which keeps track of measurements,
tests and assessment of the subject. Data collection is repeated for the subject after six
months. Subjects are followed for a maximum 48 months. During this time they track the

changes of the subject's brain structure, function, and activities that might help in understanding the progression of MCI to AD. Systematic analysis of the data provides an opportunity to compare the subjects which are in different cognitive categories.

The data set we use consists of visits of 767 subjects whose mental status diagnoses are in three different classes (NL, MCI, and AD). The physicians use laboratory investigations, cognitive testing, physical and historical data to classify the subject's mental status as NL, MCI or AD.

ADNI database contains multiple tables that contain information on subsets of clinical data. The ADNI clinical dataset is comprised of clinical information about each subject including recruitment, demographic, physical and neurological examination, cognitive assessments, patient medical history and baseline diagnosis and symptoms. All subjects have clinical and cognitive assessment and 1.5T structural MRI at 6 or 12 month intervals for 2 to 3 years. The first part of the experiment was to select the important data which plays important role in diagnosis. At the beginning through external reference we selected 118 attributes; Table 1 summarizes the list of tables from which the attributes are extracted. Our focus was on numeric data. The subject's records for selected attributes are extracted from the ADNI dataset.

Table 1: List of Table Names in ADNI Database

S #	Table Name
1	Cognitive Behavior
2	ADAS Sub-Scores and Total Scores
3	Diagnosis and Symptoms Checklist
4	Diagnostic Summary - Baseline Changes
5	Clinical Dementia Rating
6	Functional Assessment Questionnaire
7	Homocysteine – Results
8	Mini Mental State Exam
9	Neuropsychological Battery
10	Neuropsychiatric Inventory Q
11	Vital Signs

2.2   MACHINE LEARNING APPROACHES USED

In this section, the different machine learning approaches used in this study are discussed in detail.

2.2.1   Support Vector Machine (SVM)

A support vector machine is a powerful mathematical tool that performs binary classification by constructing a hyper-plane that optimally separates the data into two categories [11]. To calculate the margin, two parallel hyper-planes are constructed, one on each side of the separating hyper-plane. A good separation is achieved by the hyper-plane that has the largest distance to the neighboring data points of both classes. In the case of linearly separable data once the optimal separating hyper-plane is found, the data points closest to it are known as support vectors. The separating hyper-plane is a linear combination of those support vectors. The object of SVM is to find an optimal hyper-plane and optimize it so that all the data should be easily separated and classified correctly. This hyper-plane is not unique and it can be estimated in a way that maximizes

the performance of the classifier. The support vectors that are selected are usually small in number. The number of features in the training data does not affect the complexity of the SVM and therefore SVMs are well suited for the datasets that have a large number of features. It is not always possible to linearly separate the data. One solution is to map the data onto a higher dimensional space and define the hyper-plane there. This higher dimensional space is called the transformed feature space. In the transformed feature space any data can be made separable. The SVM uses a Kernel function to map the data into feature space where they can be linearly separable. Figure 1 shows a typical SVM.
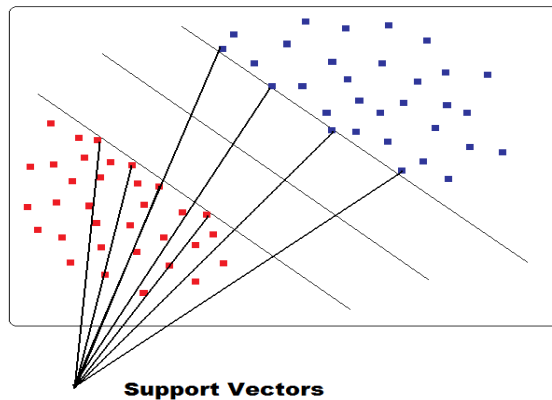
Figure 1: Support Vector Machine

2.2.2   Random Forests (RF)

A random forest is a kind of ensemble classifier composed of tree-structured classifiers [12]. It can be used for classification or regression. To grow these ensembles, often random vectors are generated that govern the growth of each tree in the ensemble. To classify a new instance represented by an input vector, each tree in the forest votes for that class and classification with the most votes is selected. All trees in the forest have the same distribution. If the number of cases in the training set is **N**, a sample of **N** cases are

selected at random - but with replacement, from the original data. This sample will be the training set for growing the tree. If there are **M** input variables, a number **m<<M** is specified such that at each node, **m** variables are selected at random out of the **M** and the best split on these **m** is used to split the node. The value of **m** is held constant during the forest growing. Each tree is grown to the largest extent possible. There is no pruning. Over fitting is not a problem in RFs. Furthermore, RFs are not sensitive to the outliers in training data.

2.2.3   Artificial Neural Networks (ANN)

A neural network is composed of interconnected units called neurons that co-operate to perform desired functions. Interconnections are used for sending signals among neurons and have weights associated with them. The output of a neuron is a weighted sum of the inputs plus a bias. The output of the entire neural network is based on the computation of all the outputs of neurons. Single layered Perceptron and multilayered Perceptrons are the most common neural networks. The multilayered Perceptron neurons are arranged into an input layer, which receives information to be processed, output layer, where the results of the processing are found and one or more hidden layers. In feed-forward neural networks, the network signals travel one way from input to output.
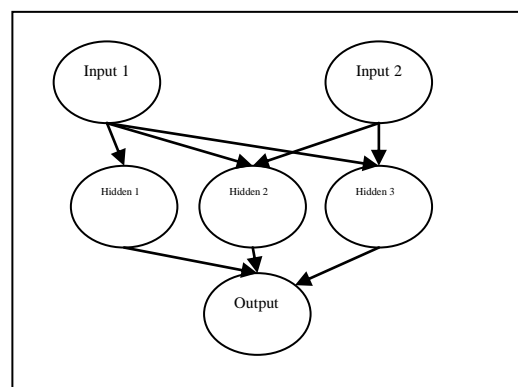
Figure 2: Neural Network

The network is trained on a set of paired data to determine input-output mapping. Once the weights of the neurons are adjusted the network is used for the classification of new data. The values of the input on an instance are presented to the input units. The output of the network is compared with the desired output and the weights are adjusted so that the network's output comes closer to the desired output. Back propagation (BP) is one of the algorithms that are used to train the networks. BP algorithms are too slow for some applications as they have to perform number of weight modifications before it reaches to a good weight configuration. A typical ANN is shown in Figure 2.

2.2.4   Bayesian Network (BN)

A Bayesian Network is a graphical model for probability relationship among a set of features [13]. It is a directed acyclic graph of nodes and links. Nodes represent features or classes, while links between nodes represent the relationship between them. Nodes are in one-to-one correspondence with features. Conditional probability tables determine the strength of the links. There is one probability table for each node (feature) that defines the probability distribution for the node given its parent nodes. If a node has no parents then its probability distribution is unconditional. If a node has one or more parents the probability distribution is a conditional distribution, where the probability of each feature value depends on the values of the parents. The most interesting feature of BNs is that they take prior information about a given problem. BNs are not feasible for problems with a very large number of features. The reason is that they would have to construct a large network which consumes time and space.

2.2.5   Simple Logistic (SL)

Simple logistic combines logistic regression and tree induction methods to classify datasets [14]. Both logistic regression and tree induction methods have their own pros and cons, which are somehow complementary to each other. The former is known by its stability in model building but with potentially high bias. The latter searches a broader range of models to capture the nonlinearity of dataset and decrease the potential bias but this, on the other hand, makes it unstable. To combine these two methods is a natural idea to get around the disadvantages by using their complementary properties. Specifically, a logistic regression model is used in the leaves of trees.

In the simple logistic model approach, LogitBoost [14] with simple regression functions as base learners is used for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection. Logistic [15] is an alternative implementation for building and using a multinomial logistic regression models with a ridge estimator to guard against over fitting by penalizing large coefficients.

2.2.6   Simple Carts (SC)

Simple Carts is a classification method implementing minimal cost-complexity pruning as a decision tree method [16]. The minimal cost-complexity pruning is to define the cost-complexity measure Ra(T) as :

 Ra(T) = R(T) + **a**| T' |

Where | T' | is the number of terminal nodes of the tree, and R(T) is the misclassification cost of the tree. So Ra(T) is formed by adding to the misclassification cost of the tree a

cost penalty for complexity [16]. The goal is to find a sub-tree T(a) to minimize Ra(T). By using different values of the constant **a**, different penalty weights can be assigned to trees with large size. Specifically, if **a** is small, | T' | could become large because it suffers small penalty for its large size. On the other hand, if **a** is large, then the tree size will have to be small to reduce the effect of large penalty. Furthermore, when **a** is extremely large, only the root node will be retained, and thus the tree will have been completely pruned.

2.2.7 Bootstrap aggregating

Bootstrap aggregating (Bagging) [25] is a meta-learner that makes decisions from multiple classifiers. Bagging generates random samples of the given training data where the sampling is done with replacement - i.e. some of the examples may be repeated in each sample. Each sample of this kind is called a bootstrap. A classifier is trained using each sample. The final classification is made based on voting by the classifiers resulting from training on each bootstrap. Bagging is used to improve the classification accuracy and stability of classifiers. It can be used for both classification and regression. For regression the final output is the average of the results of each bootstrap. Bagging is found to be useful for the classifiers that are unstable and vary the results when a small change is made in the training dataset.

2.2.8 Boosting

Boosting is a machine learning method for performing supervised learning [27]. The underlying idea is to boost the weaker learners to stronger learners. A weak learner is defined to be a classifier whose predictions are only slightly correlated with the true classification. In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification. In boosting the weak learner is used repeatedly on modified

versions of the data and combines the resulting hypotheses to get better results. In some cases boosting performs better than bagging. Boosting will not work well with noisy data because then there will be more weight to the misclassification. Adaboost [27] is an example of a boosting method.

2.2.9 Stacking

Stacking is another approach for Meta-classification [27]. It uses the outputs of more than one classifier to get an optimal result. Unlike bagging and boosting, Stacking combines models of different types. In stacking there are two levels of classification. The first is base classifiers and the second is Meta classifier. Base classifiers are run on the training dataset and their outputs are combined using the Meta classifier. Same or different classifiers can be used in both levels. Stacking takes the best from everything but it is hard to chose the combination of classifiers on both levels to get optimal results.

2.2.10 Feature Subset Selection Algorithms

Feature selection, is the technique of selecting a subset of relevant features for building robust learning models. By removing most irrelevant and redundant features from the data, feature selection helps improve the performance of learning models. Feature selection methods results in speeding up the learning process. Feature selection also helps researchers in understanding the data by telling them which are the important features.

Feature subset selection algorithms are divided into the following groups:

 Wrapper Models use a search algorithm to search through the space of possible features and evaluate each subset by building a model using only the features in that subset. The quality of the subset is usually determined using a testing set or cross-validation. Wrapper

models are widely used but they are computationally more expensive than the filter models. In the wrapper method, the machine learning algorithm is used in conjunction with the feature subset selection process.

Filter Models are similar to Wrappers in the search approach, but instead of evaluating by learning a model, a simpler filter performance measure is evaluated. Filters are less time consuming than wrapper models. In the filter approach irrelevant features are filtered out and the relevant data is presented to the machine learning algorithm.

Chapter 3

IMPROVING COMPUTER AIDED DIAGNOSIS OF ALZHEIMER'S DISEASE
WITH MACHINE LEARNING

3.1   INTRODUCTION

In this part of the thesis we study the performance of many machine learning methods on clinical dataset that is extracted from Alzheimer's disease Neuroimaging Initiative (ADNI) database. The details of the dataset are discussed in Chapter 2. We used 118 features for classification and then we used feature subset selection algorithms to reduce the number of features. We select the best six classifiers based on accuracy which are Simple Logistic (SL), Support vector Machine (SVM), Random Forests (RF), Neural Networks (NNet), Bayesian Net (BNet) and Simple cart. The accuracy results of other classifiers are also presented. We found that the level of classification accuracy is improved using reduced features. Our classification results are promising in comparison to literature reports [6].

3.2   CLASSIFICATION DATASET

In the ADNI dataset we have multiple records for the same subject as there were multiple visits of subjects. Each subject was diagnosed in each interval of 6 months as MCI, normal or AD. We made a new data set based on intervals; for example the month six dataset contains the entire subjects' data and their diagnoses in month six. The month

twelve dataset contains the subjects' data at month 12 and their diagnoses. In same way we made month twenty four and thirty six data set. Our focus is to correctly classify the subject's diagnosis as MCI, normal or AD using machine learning methods for each of the visits (month six, month twelve, month twenty four, month thirty six). Table 2 shows a brief description of the datasets.

Table 2: Classification Dataset

Data set	Instances	Normal	MCI	AD	Features	Classes
Month06	767	225	341	201	118	3
Month12	688	195	287	206	118	3
Month24	587	241	181	192	118	3
Month36	47	30	8	9	118	3

3.3 EXPERIMENTAL SETUP

We use the University of Waikato's WEKA [20] software package, an open source data mining software written in Java, to run the experiments. The Weka package has implementations of various popular machine learning algorithms. In the current study we use multiple machine learning methods and compare their accuracy for automatic classification of the cognitive status of a subject as normal, MCI or AD. We have four types of datasets; each dataset is the diagnosis of the subject in different stages. First, the classification methods were applied to all datasets without performing any feature selection. To evaluate the robustness of the classifier, the normal methodology is to perform cross validation. Cross validation is a method to estimate the error rate in a fair and unbiased way. Ten fold cross validation has been shown to be statistically adequate in evaluating the performance of a classifier [21]. In ten-fold cross validation, the training

set is equally divided into 10 different subsets. Nine out of ten of the subsets are used to train the learner and the tenth subset is used as the testing set. The procedure is repeated ten times, with a different subset being used as the testing set. For consistency, the same data were used to train and test the classifiers. The second part of the experiment is to apply feature subset selection algorithms to the dataset to find the most relevant features and eliminate the redundant or useless features that may confuse classifiers.

3.4   EVALUATIONS

In this study we applied classification and feature selection methods to all four datasets described in Table 2.

3.4.1   Pre-processing

We performed two steps of pre-processing. First we replaced all the missing values with the mean by using a method called '*ReplaceMissingValues*' in the Weka package [20]. In the second step we converted some attributes from numeric values to nominal, which are supposed to be considered categorical. For Bayesian classifiers we discretized all the attributes.

3.4.2   Classification

We applied 38 classification methods to all datasets. Details of the classification results are described in the next section. Classification methods are divided into five different categories, namely Bayesian classifiers, Function classifiers (Neural network, Regression, logistic etc), Rule based classifiers, Tree classifiers, and Lazy learner classifiers. Names of the classifiers describe the classification methods. In Chapter 2 we discussed the details of only selected methods that overall performed better than other classifiers.

3.5   EVALUATION METRICS

 A classification method's performance is evaluated using Precision, recall and accuracy.

Precision: is the proportion of the examples which truly have class x among all those

which were classified as class x. ***Recall:*** is the proportion of examples which were

classified as class x, among all examples which truly have class x. ***Accuracy:*** is the

percentage of correctly classified instances over the total number of instances

3.6   RESULTS

We compare the performance of many classifiers over all datasets. Tables 3 through 6

show the accuracy of the selected methods which performed well in all datasets. In the

figures we give all the classifiers' accuracy results in the form of a graph. Over all we

used 38 methods for classification without performing feature selection.

Table 3 shows the accuracy results for the month 06 dataset. The maximum accuracy

achieved for the month 06 dataset is 88.7875 by the simple logistic method. For month 12

and 24 datasets FURIA method outperformed all other classifiers attaining a maximum

accuracy of 86.7733   and 88.4157 respectively. For the month 36 dataset the maximum

accuracy is 93.617 by AODEsr classifier, which is quite promising in comparison to

literature report [6]. As shown in Table 2, the month 36 dataset is small as compared to

the other datasets. However, we used 10 fold cross validation to reduce the size effect.

Table 3: Month 06 Dataset Classification Results

Classifier	Accuracy	Precision			Recall		
		NL	MCI	AD	NL	MCI	AD
MLP	85.3977	0.929	0.825	0.82	0.924	0.856	0.771
Simple Logistic	**88.7875**	0.955	0.864	0.853	0.951	0.891	0.811
SMO	85.9192	0.939	0.842	0.795	0.96	0.845	0.771
Random Forest	86.9622	0.022	0.15	0.042	0.947	0.827	0.86
Simple Cart	87.3533	0.955	0.828	0.865	0.951	0.906	0.731

Table 4: Month 12 Dataset Classification Results

Classifier	Accuracy	Precision			Recall		
		NL	MCI	AD	NL	MCI	AD
MLP	78.6337	0.856	0.734	0.787	0.881	0.739	0.757
Simple Logistic	84.157	0.925	0.802	0.813	0.921	0.808	0.81
SMO	82.8488	0.899	0.779	0.827	0.921	0.804	0.771
Random Forest	81.8314	0.911	0.753	0.823	0.916	0.815	0.729
Simple Cart	85.7558	0.95	0.828	0.809	0.941	0.819	0.829
Bayes Net	80.5233	0.819	0.777	0.825	0.941	0.721	0.786
FURIA	**86.7733**	0.936	0.828	0.856	0.936	0.855	0.819

Table 5: Month 24 Dataset Classification Results

Classifier	Accuracy	Precision			Recall		
		NL	MCI	AD	NL	MCI	AD
MLP	80.4089	0.823	0.665	0.9	0.865	0.673	0.852
Simple Logistic	87.2232	0.928	0.789	0.885	0.938	0.766	0.897
SMO	82.9642	0.852	0.735	0.874	0.927	0.667	0.87
Random Forest	85.5196	0.902	0.744	0.903	0.907	0.766	0.879
Simple Cart	84.1567	0.951	0.714	0.855	0.912	0.76	0.843
Bayes Net	84.4974	0.851	0.747	0.914	0.948	0.708	0.861
FURIA	**88.4157**	0.932	0.817	0.892	0.927	0.784	0.924

Table 6: Month 36 Dataset Classification Results

| Classifier	Accuracy	Precision			Recall		
		NL	MCI	AD	NL	MCI	AD
MLP	87.234	0.906	0.667	0.889	0.967	0.5	0.889
Simple Logistic	85.1064	0.857	0.667	0.889	1	0.25	0.889
SMO	85.1064	0.857	0.667	0.889	1	0.25	0.889
Random Forest	80.8511	0.857	0.4	0.857	1	0.25	0.667
Simple Cart	80.8511	0.158	0.968	0.333	0.538	1	0.125
Bayes Net	89.3617	0.909	0.8	0.889	1	0.5	0.889
AODE sr	**93.617**	0.968	0.857	0.889	1	0.75	0.889
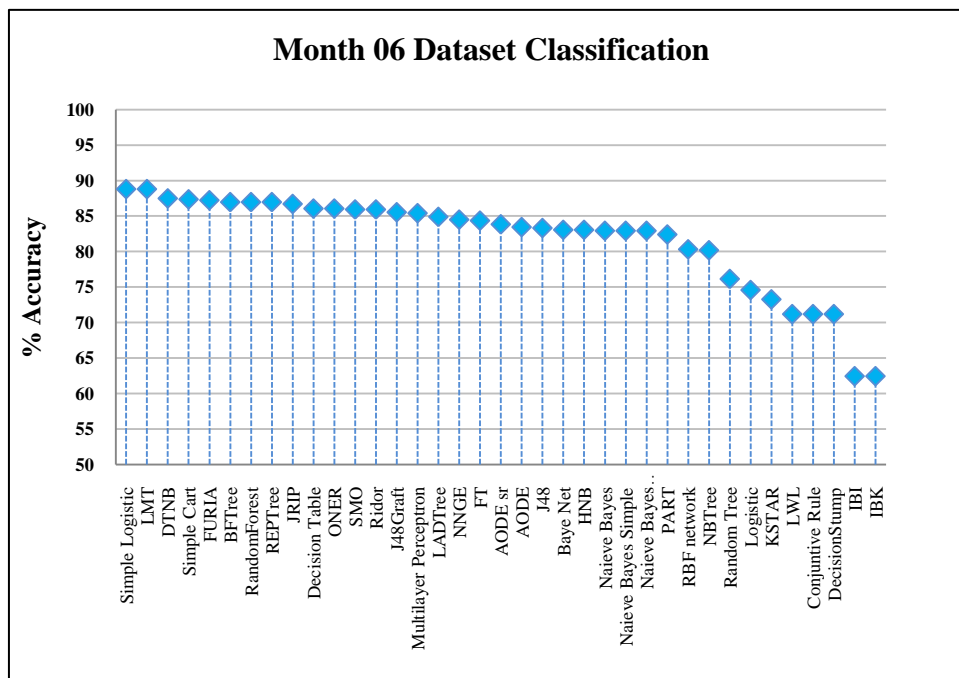
Figure 3: Performance of Classifiers on MONTH 06 dataset
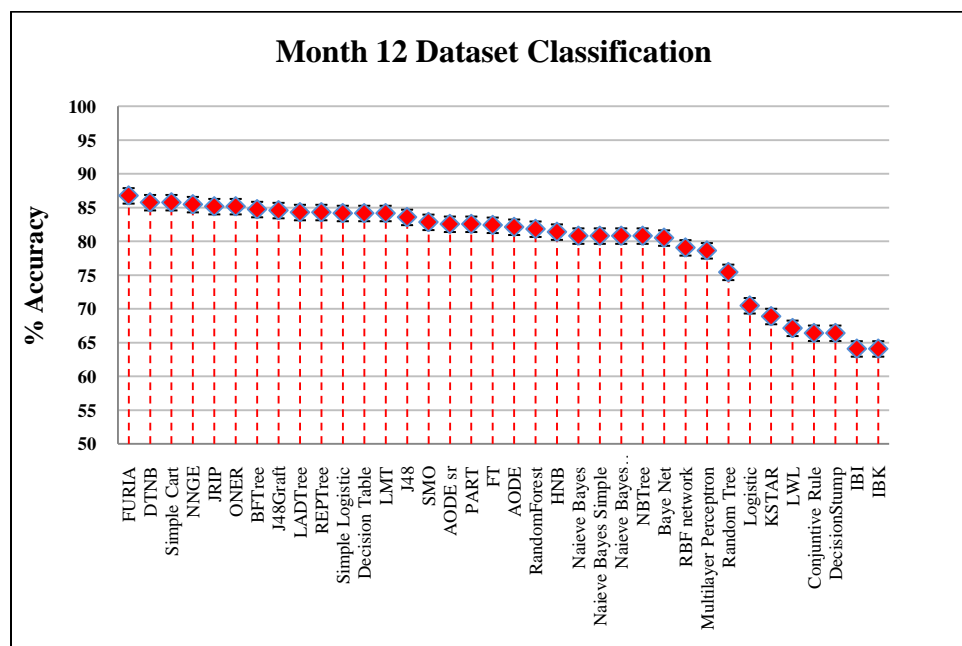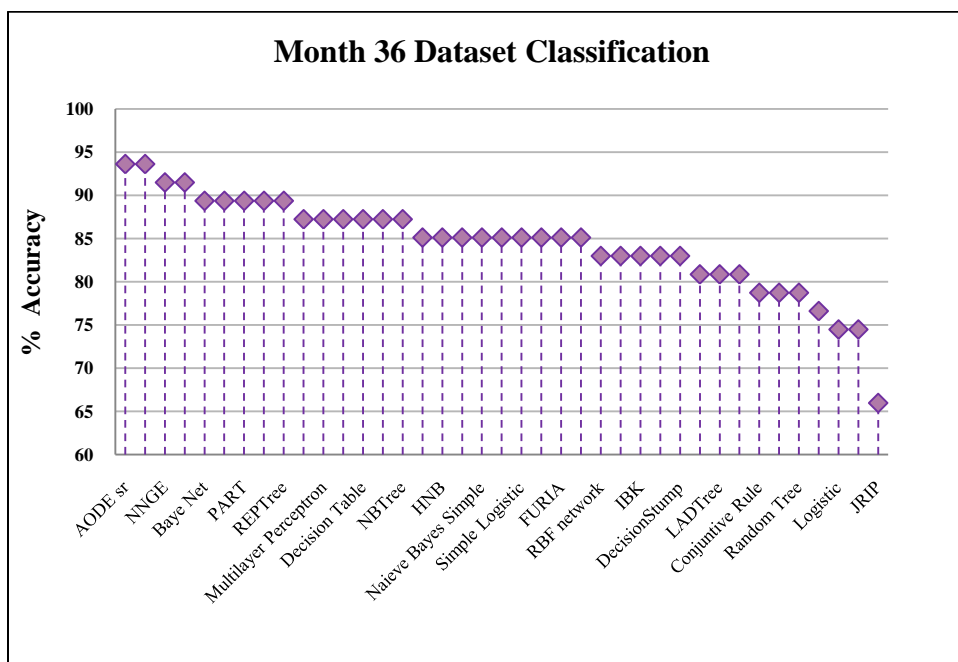

Figure 4: Performance of Classifiers on MONTH 12 dataset
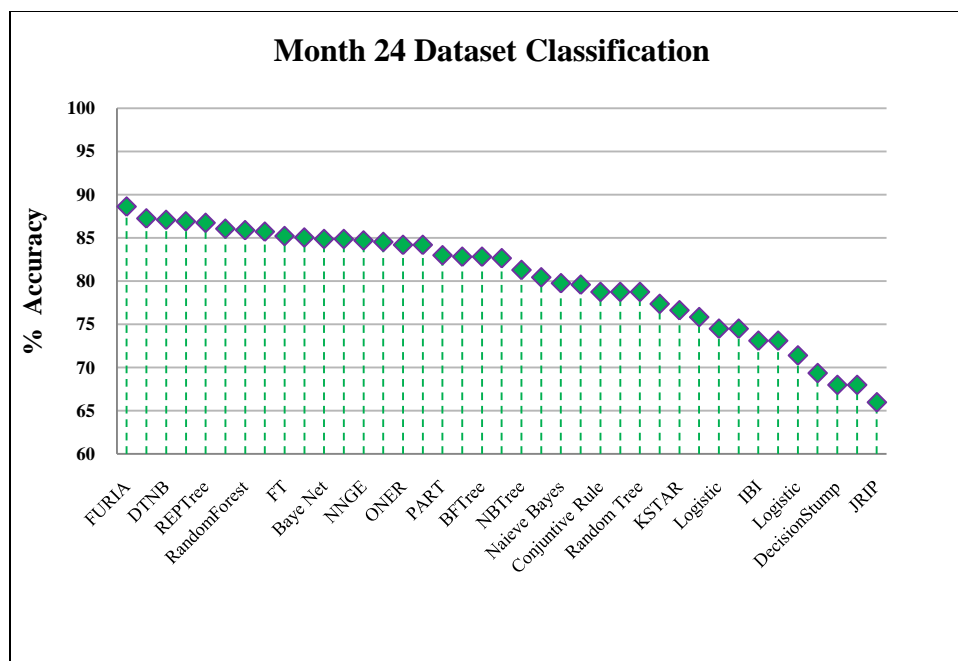
Figure 5: Performance of Classifiers on MONTH 24 dataset

Figure 6: Performance of Classifiers on MONTH 36 dataset

3.6.1   Feature Selection Results

Since the dimensionality in our dataset is very large with 118 features, our second objective was to determine the minimum number of features where the accuracy reaches maximum. We studied several feature selection methods available in Weka and found that 'wrappersubseteval' with classifier 'logistic' gives more predicative feature subsets than others. Figure 7 confirms that accuracy has been improved after applying feature subset selection method over all datasets. As far as the true positive rate (Recall) of AD is concerned, the selected features work better for Month24 and Month36 and not as good for Month06 and Month12.

The maximum number of features selected is 13 which are far fewer than 118. Table 1 shows the list of tables in ADNI database from where the attributes are selected. According to the features selected, the following tables are playing an important role in diagnosis, which are very relevant to the criteria used in clinical diagnosis [6].

1) Functional Assessment questionnaire.
2) Mini mental state exam
3) ADAS- Cognitive Behavior
4) Clinical Dementia Rating
5) Neuropsychological Battery
6) Diagnosis and Symptoms Checklist

Table 7: Accuracy of Classifiers with Feature subset selection algorithms

Dataset	Features	Classifier	Accuracy	Precision		
				NL	MCI	AD
Month 06	13	Logistic	90.09	0.956	0.912	0.821
Month 12	10	Simple Logistic	88.66	0.936	0.88	0.848
Month 24	9	Logistic	90.63	0.953	0.83	0.924
Month 36	4	Logistic	97.87	1	0.875	1
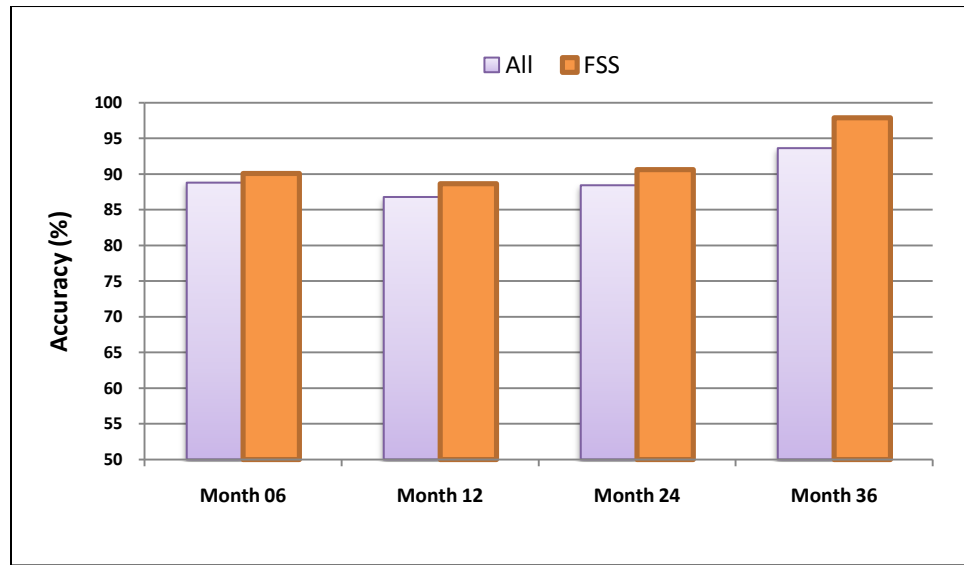
Figure 7: Performance improved with reduced features

3.6.2 Meta learner results

In order to further improve the performance of learning algorithms we used ensemble methods like bagging, stacking and boosting and compared their performance with 10-fold cross validation. We used Simple logistic, SMO, Baye Net, Simple Cart and Random forest. For stacking we chose Simple cart, Simple logistic, and random forest as base learners and the method listed at the beginning of each raw as the Meta-learner. Tables 8 through 11 show the performance comparison of Meta learners with 10 fold cross validation. The results show that in some cases Meta learners perform better than simple

learners but could not reach the maximum accuracy that is achieved by feature subset selection methods.

Table 8: Meta learners on Month 06 Dataset

Classifier	10-Fold	Bagging	Boosting	Stacking
Simple Logistic	**88.78%**	86.44%	86.04%	87.87%
SMO	85.91%	85.65%	81.35%	87.74%
Random Forest	87.35 %	88.52%	87.35%	86.31%
Simple Cart	87.35%	87.09%	88.26%	86.96%
Bayes Net	82.79 %	82.79%	82.79%	87.87%

Table 9: Meta learners on Month 12 Dataset

Classifier	10-Fold	Bagging	Boosting	Stacking
Simple Logistic	84.15%	84.59%	82.84%	**87.64%**
SMO	82.84%	83.13%	77.03%	86.62%
Random Forest	81.83%	85.46%	83.28%	86.04%
Simple Cart	85.75%	86.48%	85.61%	85.31%
Bayes Net	80.52%	80.52%	79.65%	**87.64 %**

Table 10: Meta learners on Month 24 Dataset

Classifier	10-Fold	Bagging	Boosting	Stacking
Simple Logistic	87.22%	87.05 %	86.54%	88.41%
SMO	82.96%	82.28 %	81.43%	87.56%
Random Forest	85.51%	**89.09%**	86.54%	86.37%
Simple Cart	84.15%	86.88 %	88.58%	87.39%
Bayes Net	84.49%	84.49%	84.83%	88.07%

Table 11: Meta learners on Month 36 Dataset

Classifier	10-Fold	Bagging	Boosting	Stacking
Simple Logistic	85.10%	91.48%	85.10%	87.23%
SMO	85.10%	82.97%	85.10%	91.48%
Random Forest	80.85%	87.23%	87.23%	89.36 %
Simple Cart	80.85%	87.23%	89.36%	85.10 %
Bayes Net	89.36%	**93.61%**	91.48 %	89.36 %

Chapter 4

PREDICTION OF POSSIBLE CONVERSION FROM MCI TO AD USING

MACHINE LEARNING

4.1   INTRODUCTION

There is an active research going on to delay the onset or slow down the progression of AD. Early diagnosis of AD helps both the patients and their caregivers to improve the quality of their lives. There are treatments that slow down the disease progression and help in prevention [1]. However this is only possible if the AD is diagnosed with high accuracy in its early symptomatic stage.

Many scientists believe that there is a transitional stage between normal aging and AD termed as mild cognitive impairment (MCI). During this stage a person experiences more memory loss that cannot be linked to age problems but also not severe enough to point to probable AD. MCI has high chance to turn to AD. Research shows that there is evidence that 10% to 15% of MCI subjects turn to probable AD per year. While for a healthy person it is just 1% to 2%. As a result MCI got more attention among researchers.

One of the major objectives of ADNI (Alzheimer's Disease Neuroimaging Initiative) is to find bioglocial biomarkers to measure the progression of MCI and probable AD. For that they are using advanced brain imaging techniques and clinical and neurological

assessments to assess the progression of the disease. Finding biomarkers will help in the early diagnosis of AD and in the development of treatments.

MCI does not fulfill the criteria for probable AD that makes it more challenging to predict the prior conversion from MCI to probable AD. Most of the researchers are focusing on brain imaging techniques to predict the conversion of MCI to probable AD.

The goal of our study is to predict the probable conversion of MCI to AD based on only clinical data using machine learning methods.

4.2   PREDICTION DATASET

Our prediction dataset is the same as that used for the classification problem explained in detail in the previous chapter. The data set used for this study is comprised of clinical information about each subject including recruitment, demographic, physical and neurological examination, cognitive assessments, patient medical history and baseline diagnosis and symptoms. For the prediction dataset we have two types of datasets.

4.2.1   Training Dataset

For training classifiers we have a dataset where the subjects are diagnosed in month 06 as Normal or MCI or AD. We have a total of 732 subjects' data and their diagnosis in month 06. 212 subjects are diagnosed as Normal, 341 as MCI and 179 are diagnosed as AD. However in the training dataset we will just include the data for subjects that are diagnosed as normal and AD, excluding the MCI subjects. We name this the *Baseline* dataset.

4.2.2   Testing Dataset

For testing purposes we have three datasets each dataset is the follow up diagnosis of the subject for three years after baseline. Each dataset is comprised of two groups, *MCI stable:* All subjects that were diagnosed MCI in baseline and remained MCI till three years and *MCI converters:* Subjects that were diagnosed MCI in baseline and they turned to AD in later stages. Table 12 shows the details of each group.

Table 12: Prediction Dataset Details

Baseline	Normal	MCI		AD
		MCI-S	MCI-C	
Baseline	212	NILL		179
Month 12		294	43	
Month 24		236	92	
Month 36		9	7	

4.3   EXPERIMENTAL SETUP

Our main objective in this study is to improve the prediction performance (sensitivity and specificity) by evaluating all available diagnostic information with machine learning techniques.

We again used the University of Waikato's WEKA [20] software package, to run the experiments. In the current study we use multiple machine learning methods and compare their accuracy for prediction of subjects that are diagnosed as MCI in baseline, predicting either they will remain MCI or will turn to probable AD. We used the training dataset that is already discussed in previous section-termed as *baseline*. For testing purposes we have three datasets that are the diagnoses of the subjects after baseline - i.e. month 12,

month 24 and month 36, each of these datasets are representing the diagnosis at different stages.

We train the classifier using the baseline dataset that is comprised of the subjects diagnosed as purely healthy and pure AD and test it with follow up diagnosis of MCI subjects after the baseline, where we have subjects from two groups MCI stable (those who will remain MCI) and MCI converted (those who turned to AD after baseline). We also include the subjects that were MCI in base line and turned to Normal in MCI stable. In both datasets we have two classes: 1 represents that the subject will not turn to AD and 0 represents that the subject will turn to AD. Table 13 shows an example of a confusion matrix that we will get after prediction.

Table 13: Example of Confusion Matrix

		Predicted Class	
		AD	NOT AD
Actual Values	AD	**31(TP)**	0(FN)
	NOT AD	41(FP)	**100(TN)**

In the confusion matrix rows show the actual diagnosis and columns show the predicted values after classification. So the diagonal values are the ones that are correctly classified. In the above confusion matrix 31 patients were correctly predicted to turn to AD. 41 subjects are misclassified as positive class (i.e. Turn to AD) when they actually belong to the negative class (i.e. Not turn to AD). In our case following are the definitions for true positive and true negative class.

True Positive (TP): True positives are the subjects predicted to turn to AD and actually turned to AD.

False Positive (FP): False positives are the subjects predicted to turn to AD but actually did not turn to AD and remained MCI or turned back to Normal.

True Negative (TN): True negatives are the subjects predicted not to turn to AD and actually did not turn to AD and remained MCI or turned to Normal.

False Negative (FN): False negatives are the subjects who actually did turn to AD but were predicted not to turn to AD.

For the prediction part we used multiple experimental settings. In the first part we used several machine learning approaches without using cost sensitive classifiers, while in the second part we used the same classifiers but using different cost matrix.

4.4   PERFORMANCE MEASUREMENT

Our main focus of this study is to measure the classifier's performance to predict the probable conversion from MCI to AD. It is a binary classification problem but we are more interested in one class (i.e. to predict probable AD conversion from MCI). In such cases overall classification accuracy is not the most important measure of performance. We use metrics such as Sensitivity and Specificity and the Area Under the Curve (AUC) to evaluate the performance of learning algorithms. All of those metrics are functions of the confusion matrix. The following are the definitions of the above metrics.

Sensitivity measures the proportion of actual positives which are correctly identified as such. In our case the percentage of MCI-Converted people who are correctly predicted to convert to AD. In the above example the sensitivity is 100% because all subjects of group MCI-converted are correctly predicted and there is 0 false negatives. Mathematically we can define sensitivity as:

Sensitivity = Number of True positives / (Number of True positives + Number of false Negatives)

Specificity measures the ability of the classifier to identify True negative. In our case it is the percentage of true predictions of NOT AD class. In the above example, the precision of the classifier is almost 71%. Mathematically, Specificity is defined as:

Specificity= Number of True Negatives / (Number of True negatives + Number of False positives)

Area under Curve (AUC)

To find an optimal classifier, another approach is to draw the receiver operating characteristics (ROC) curve [26]. It plots the values of two statistical measures, sensitivity and (1-specificity) computed from the results of a binary classifier system as its discrimination threshold is varied.

A completely random predictor would generate a straight line at an angle of 45 degrees with the horizontal, from bottom left to top right. Classifiers with ROC curves higher than this straight line are better than a random classifier. The statistic that is most commonly calculated from the ROC for comparing classifier performance is the Area under ROC Curve (AUC).
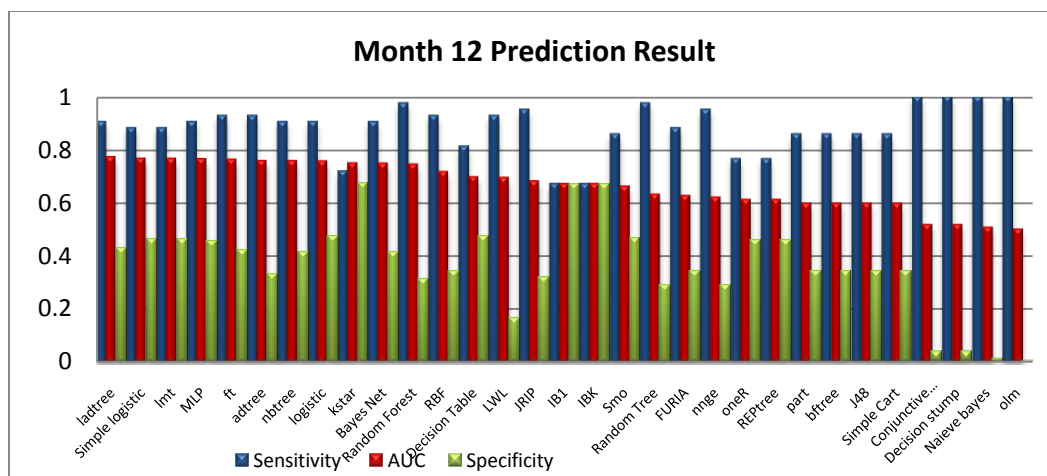
4.5   PREDICTION RESULT

Figures 8 through 10 show the prediction accuracy of different machine learning algorithms based on sensitivity, specificity and AUC. In the figures, the learning algorithms are sorted in descending order of AUC.

The results show that for all classifiers there is a tradeoff between sensitivity and specificity. In our case our testing class is imbalanced and the rare class is of more interest. Thus it is desirable to have a classifier that gives high prediction accuracy over the rare class (AD), while keeping reasonable accuracy for the majority class (NOT AD). We used all of these metrics to get a classifier that gives best result.

As we discussed earlier, overall accuracy is not important in the current scenario. Prediction results show that all classifiers' sensitivity is higher than specificity for all datasets.

LADTree, ADTree and FT classifiers are selected as best classifiers based on AUC for Month 12, Month 24 and Month 36 accordingly. We can see that there are classifiers with higher sensitivity (i.e. 100%) but their specificity values are very low and so are their AUCs. Cost sensitive classifiers are used to manage this trade-off between sensitivity and specificity as discussed in the next section.

Figure 8: Month 12 Prediction accuracy comparison
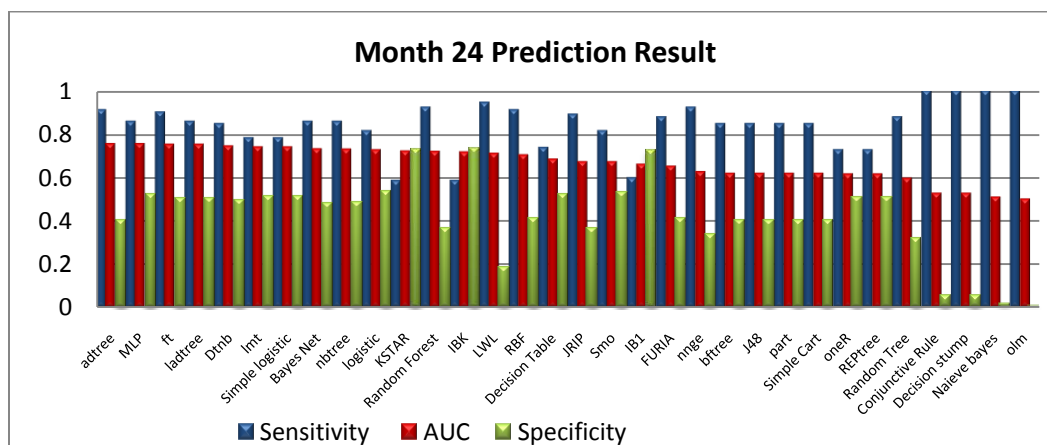
Figure 9: Month 24 Prediction accuracy comparison
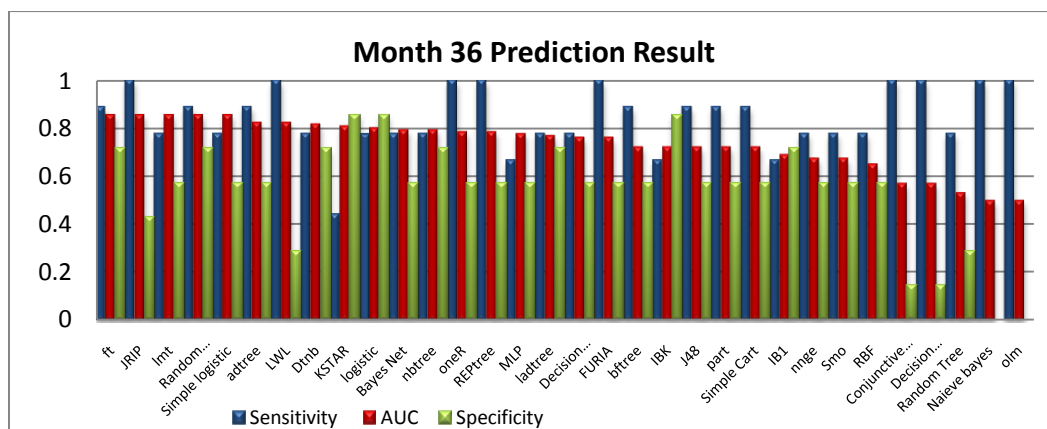
Figure 10: Month 36 Prediction accuracy comparison

4.5.1   Cost Sensitive Classification Results

Machine learning algorithms are typically tuned to minimize the classification error. In

the current situation we are not as interested in the overall classification accuracy. In fact

sensitivity and specificity are more important. For each classifier there is a cost

associated with each misclassification that a classifier makes. By default the value of that

cost is equal for both classes. The misclassification cost can be changed to make the

classifier more biased for the desired class. In our case we are already getting more than

80% accuracy for the class AD (i.e. sensitivity). One way to maximize the specificity is

to increase the cost for misclassification of class NOT AD. In the second part, we

generalized all the algorithms to take into account the misclassification costs for NOT

AD class. The misclassification cost is a function of the predicted class and the actual

class and can be represented as a matrix with misclassification cost associated with false

positive and false negative class. This cost matrix is an additional input to the classifier to

take into account the importance of the desired class.

Table 14: Example of Cost Matrix

		Predicted Class	
		AD	NOT AD
Actual Values	AD	**0**	1
	NOT AD	1	**0**

By default the misclassification cost for both classes are the same (i.e. 1) as shown in

Table 14.

In second part of the prediction experiment, we used different costs to optimize the

learning algorithm. For this report we only included the one that maximized the

specificity and also did not harm sensitivity of the classifiers. We used the same classifiers discussed in the previous section but this time with the following cost matrix:

Table 15: Cost Matrix for Meta cost classifiers

		Predicted Class	
		AD	NOT AD
Actual Values	AD	0	1
	NOT AD	**10**	0

Figures 11 through 13 show the comparison of Meta cost classifiers on all datasets. The results show that after using the cost matrix, for almost all classifiers, specificity increased but this also results in decreasing the sensitivity by about 1 to 2%.

Table 16 shows the best classifiers for prediction of MCI to probable AD conversion. The selection is not based on overall accuracy but instead based on the AUC.

Table 16: Best classifiers based on AUC

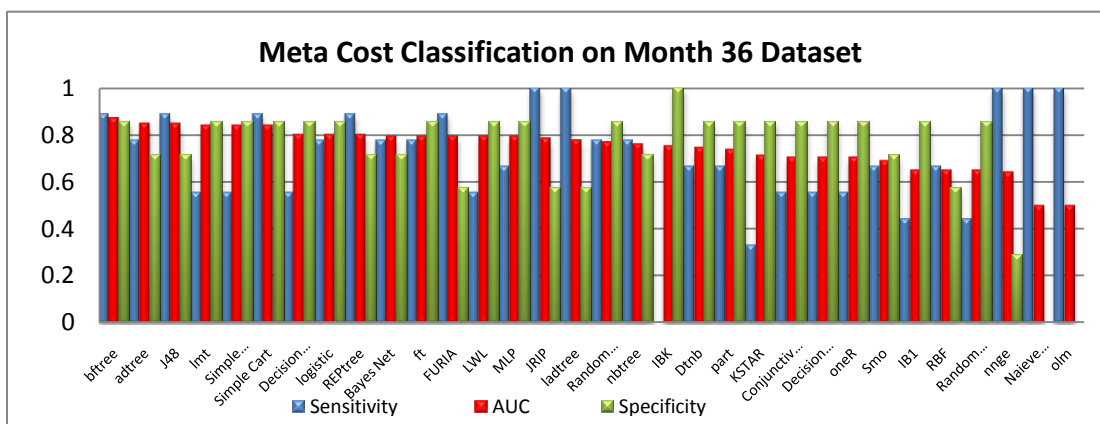| Dataset	Classifier	Accuracy	Sensitivity	Specificity	AUC
Month 12	LMT	69.73 %	0.814	0.68	0.786
Month 24	ADTree	54.57%	0.913	0.403	0.758
Month 36	BFTree	87.50%	0.889	0.857	0.873
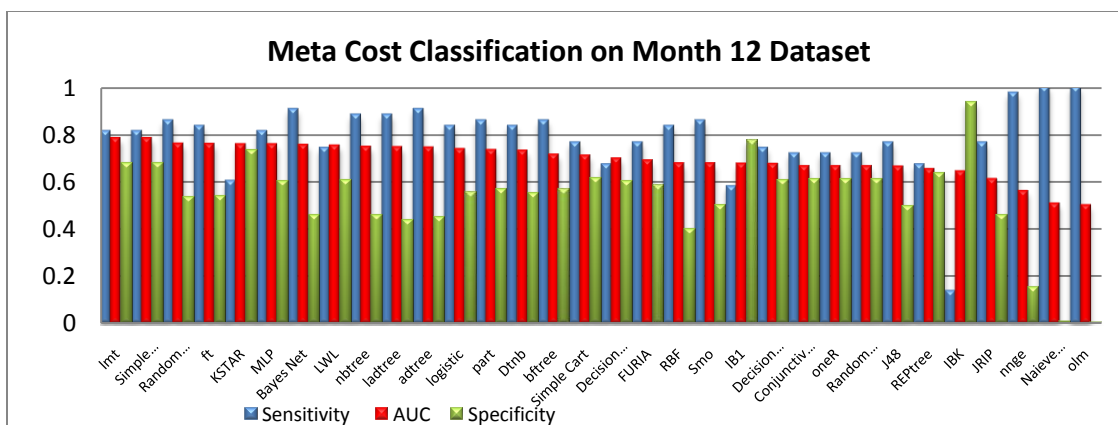
Figure 11: Meta cost classifiers accuracy comparison on Month 12 Dataset
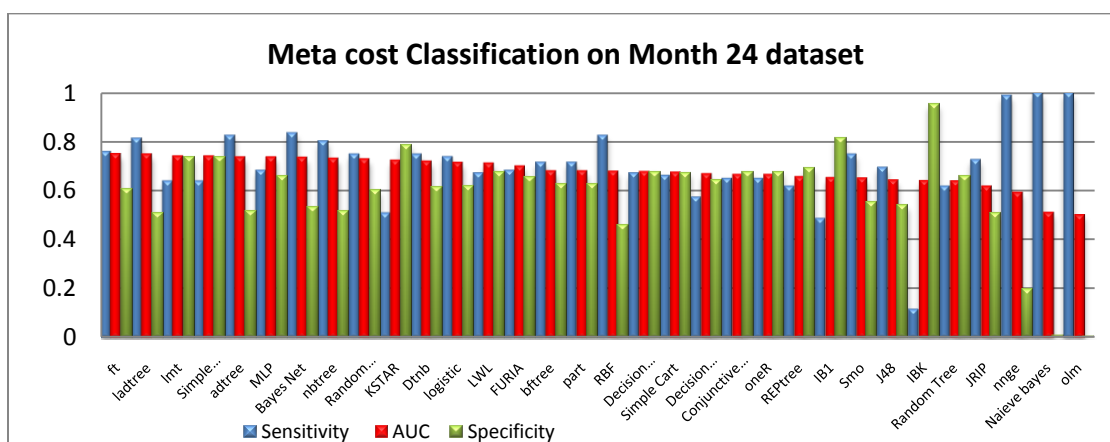
Figure 12: Meta cost classifiers accuracy comparison on Month 24 Dataset

Figure 13: Meta cost classifiers accuracy comparison on Month 36 Dataset

Chapter 5

CONCLUSION

This study applies a variety of models for the automatic classification of Alzheimer's disease, MCI and Normal subjects based on different cognitive tests, physical examinations, age, mental status examination and neuropsychiatry assessments. The results of this study can help in improving the diagnosis performance by physicians. Results show that we can improve the diagnosis performance of Alzheimer's diseases using machine learning methods where the maximum accuracy is more than 85%. Furthermore, with the help of feature selection methods, more than 90% accuracy can be achieved. These results can help in the early diagnosis of AD. For the prediction of MCI to probable AD conversion, our results are promising. We get up to 87% sensitivity but there is a trade-off between sensitivity and specificity which can be controlled using cost-sensitive classifiers. Considering sensitivity, specificity and AUC we still get very promising results. These results are very helpful for predicting the future diagnosis of MCI patients that either they will remain MCI or will convert to AD. However, we cannot directly compare our results to the few other published papers in this domain as our datasets and methods are different.

5.1 FUTURE WORK

In the current study we only used the clinical data which included neuropsychological assessments, demographic information, physical and neurological examination, cognitive assessments, patient medical history and baseline diagnosis and symptoms. In the future we can use advanced brain images in conjunction with the clinical data to improve the diagnosis and the prediction processes.

REFERENCES

[1] Berchtold NC, Cotman CW (1998). "Evolution in the conceptualization of dementia and Alzheimer's disease: Greco Roman period to the 1960s". Neurobiol. Aging 19 (3): 173–89

[2]   Alzheimer's Association, 2010 Alzheimer's disease Facts and Figures, Alzheimer's & Dementia, Volume 6

[3]   ALZHEIMER'S DISEASE Unraveling the mystery National Institutes of Health NIH Publication Number: 02-3782 December 2003

[4]   Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI), Alzheimer's & Dementia 1 (2005) 55–66

[5]   Winblad B, Palmer K, Kivipelto M, et al. Mild cognitive impairment—beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. J Intern Med 2004; 256(3):240–6.

[6]   Fleisher AS, et al, Alzheimer's Disease Cooperative Study, Clinical predictors of progression to Alzheimer disease in amnestic mild cognitive impairment. Neurology. 68(19):1588-95. 2007

[7]   Classification and treatment of different stages of Alzheimer's disease using various machine learning methods International Journal of Bioinformatics Research, Volume 2, Issue 1, 2010, pp-44-52

[8]   A Predictive Model for Identifying Possible MCI to AD Conversions in the ADNI Database Wuhan, China Knowledge Acquisition and modeling, 2009. KAM '09, Volume 3, 2009 pages 102-105

[9]   Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI, Neuroradiology, Volume 51, Issue 2, page 73.

[10]   Early Alzheimer's disease diagnosis using partial least squares and random forests, Neuroscience Letters Volume 472, Issue 2, 19 March 2010, Pages 99-103

[11]   Corinna Cortes and V. Vapnik, "Support-Vector Networks", Machine Learning, Volume 20, November 3, 1995, page 273-297

[12]   Leo Breiman, Random Forests, Machine Learning, v.45 n.1, p.5-32, October 1 2001 .

[13]   Pearl, J, "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning" (UCLA Technical Report CSD-850017). Proceedings of the 7thConference of the Cognitive Science Society, University of California, Irvine, CA. 1985 pp. 329–334.

[14]   Niels Landwehr , Mark Hall , Eibe Frank, Logistic Model Trees, Machine Learning, v.59 n.1-2, p.161-205, May 2005

[15]   Le Cassie S, Van Houwelingen J (1992) Ridge estimators in logistic regression. Applied Statistics 41: 191–201.

[16]   Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, California.

[17]   M. A. Hall (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.

[18]   Liu, H., and Setiono, R., (1996). A probabilistic approach to feature selection - A filter solution. In 13th International Conference on Machine Learning (ICML'96), July 1996, pp. 319-327. Bari, Italy.

[19]   Ron Kohavi, George H. John (1997). Wrappers for feature subset selection. Artificial Intelligence. 97(1-2):273-324.

[20]   G. Holmes, A. Donkin, and I.H. Witten. Weka: A machine learning workbench. In Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems, 1994.

 [21]   S. M. Weiss and C. A. Kulikowski, Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning and Expert Systems. Morgan Kaufmann, San Mateo (1991).

[22] Nho K., Shen L., Kim S., Risacher S.L., West J.D., Foroud T., Jack C.R., Weiner M.W., Saykin A.J.Automatic Prediction of Conversion from Mild Cognitive Impairment to Probable Alzheimer's Disease using Structural Magnetic Resonance Imaging. AMIA Annu Symp Proc. 2010 Nov 13; 2010:542-6. PMID: 21347037. PMCID: PMC3041374

[23] Marko Robnik Sikonja, Igor Kononenko: An adaptation of Relief for attribute estimation on regression. In D.Fisher (ed.): Machine Learning, Proceedings of 14th International Conference on Machine Learning ICML'97, Nashville, TN, 1997.

[24]   Guyon, J. Weston, S. Barnhill, V. Vapnik (2002). Gene selection for cancer classification using support vector machines. Machine Learning. 46:389-422.

[25] BREIMAN, Leo, 1996. Bagging Predictors. Machine Learning, 24(2), 123–140

[26] Green DM, Swets JA. Signal Detection Theory and Psychophysics. Wiley, 1974.

[27] Ian Witten, Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques (2nd edition). Morgan Kaufmann, 2005.

.