Machine Learning Techniques for Diagnostic Differentiation of Mild Cognitive Impairment and Dementia

Jennifer A. Williams, Alyssa Weakley, Diane J. Cook, Maureen Schmitter-Edgecombe
Washington State University, Pullman, WA
{jen_williams, alymae, schmitter-e}@wsu.edu, cook@eecs.wsu.edu

Abstract

Detection of cognitive impairment, especially at the early stages, is critical. Such detection has traditionally been performed manually by one or more clinicians based on reports and test results. Machine learning algorithms offer an alternative method of detection that may provide an automated process and valuable insights into diagnosis and classification. In this paper, we explore the use of neuropsychological and demographic data to predict Clinical Dementia Rating (CDR) scores (no dementia, very mild dementia, dementia) and clinical diagnoses (cognitively healthy, mild cognitive impairment, dementia) through the implementation of four machine learning algorithms, naïve Bayes (NB), C4.5 decision tree (DT), back-propagation neural network (NN), and support vector machine (SVM). Additionally, a feature selection method for reducing the number of neuropsychological and demographic data needed to make an accurate diagnosis was investigated. The NB classifier provided the best accuracies, while the SVM classifier proved to offer some of the lowest accuracies. We also illustrate that with the use of feature selection, accuracies can be improved. The experiments reported in this paper indicate that artificial intelligence techniques can be used to automate aspects of clinical diagnosis of individuals with cognitive impairment.

Introduction

Accurate classification of cognitive impairment has benefits of personal and medical importance. In clinical settings, manual diagnosis of cognitive impairment is time intensive and can require multiple pieces of information (e.g., neuropsychological test scores, laboratory study results, knowledgeable informant reports). These data are assembled together to create a cohesive picture of the individual's impairment where efficiency and accuracy are governed by a practitioner's level of expertise. Furthermore, the monetary expense of a medical diagnosis is often a primary concern, thus making reliable alternatives to traditional medical diagnosis valuable.

In this paper, machine learning algorithms are explored to determine if the analysis of neuropsychological and demographic data can be automated for the diagnosis of cognitive impairment. In a recent study, Chen & Herkovits (2010) found that of several different statistical and machine learning approaches, a support vector machine (SVM) and a Bayesian-network classifier were the best methods for classifying participants with very mild dementia or no dementia. Shankle et al. (1998) also found success in using a Bayesian classifier to predict scores on a Clinical Dementia Rating (CDR; Morris 1998) of individuals exhibiting very mild, moderate-to-severe, or no dementia. The CDR is a two-stage process that takes approximately 60 minutes and requires interviews with both the individual with cognitive impairment and a significant other. The authors were unable to exceed the CDR's inter-rater reliability of approximately 80% (McCulla et al., 1989; Morris, 1997; Burke et al., 1998). When the authors combined mild dementia with the very mild dementia criterion, their approach achieved a classification accuracy of 95%. While increased accuracy was realized by combining two groups, it is more valuable to be able to discriminate between groups, considering very mild dementia represents a preclinical form of dementia, while mild dementia exceeds the threshold of probable dementia.

Artificial neural networks (NN) have also been utilized to classify mild cognitive impairment (MCI), characterized by cognitive deficit(s) that are greater than that considered to be "healthy aging" but fail to meet criteria for dementia (Petersen et al. 2001), dementia, and healthy control participants using neuropsychological data (Quintana et al. 2012). Using all three diagnostic groups, the NN was able to correctly classify 66.67% of the participants. When the model only included healthy controls and MCI participants, diagnostic accuracy increased to 98.33%. Finally, when the model was comprised of AD and healthy control

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

participants, the model classified all individuals correctly. While the latter two results are impressive, clinical utility is maximized when classifiers are sensitive to more than two stages of cognitive decline.

As suggested above, healthy older adults and MCI or healthy older adults and dementia have been discriminated using machine learning methods. However, differentiating these three categories using a single model has been problematic. The purpose of the present research was to use neuropsychological and demographic data to predict (1) CDR scores (i.e., normal [CDR = 0], very mild dementia [CDR = 0.5; proxy for MCI], and dementia [CDR = 1]) and (2) clinical diagnoses (i.e., cognitively healthy, MCI, dementia) through the implementation of four well-known machine learning models: NN, SVM, naïve Bayes (NB), and decision tree (DT).

In addition to classifying individuals, a secondary goal of the project was to determine a small set of attributes (i.e., neuropsychological tests, demographic information) that can be used to reliably diagnose individuals. By isolating a select and reduced number of commonly used clinical measures, it is plausible that a drastic reduction in test administration time, which ultimately translates into diagnostic cost, could be realized.

Based on prior research (Shankle et al., 1998; Chen & Herskovits 2010), SVM and Bayesian models are hypothesized to optimize our performance criterion with the fewest number of misclassified individuals and the lowest model error. Because clinical diagnoses were originally made by considering scores on neuropsychological measures, we expect the machine learning algorithms to achieve higher classification when predicting diagnostic values than predicting CDR values.

Method

Three categories of datasets were used in the current study. The first datasets used clinical diagnosis as the class that was being predicted. The first dataset will be referred to as the Diagnosis dataset. The second and third datasets used CDR scores as the class that was predicted. The second dataset is referred to as CDR dataset. The final dataset used a semi-supervised approach to provide labels for the CDR dataset containing data points with no target class value. This third dataset will be referred to as the SS-CDR dataset. As many of the same attributes were used to originally diagnose participants (in addition to collateral data) a validation problem exists when the same data is used in the machine learning models. To circumvent this problem, CDR scores were also included in the current study as a class since there was no overlap between the information used to assign CDR scores and the attributes used in the machine learning models. As fewer participants completed the CDR than participated in the study, there was a desire to increase the data available for the machine learning models. Therefore, the additional semi-supervised learning approach was employed.

Participants

Participants in the Diagnosis dataset included 53 individuals with dementia (52 participants suspected of probable AD, 1 suspected of Lewy Body Dementia), 97 individuals with MCI, and 161 cognitively healthy older adult control participants. Participants in the CDR datasets included 154 individuals that received a CDR of 0, 93 individuals, with a CDR of 0.5, and 25 individuals with a CDR of 1.0 (mild dementia) or 2.0 (moderate dementia). CDRs of 1.0 and 2.0 were combined as only 8 participants fell within the CDR = 2.0 category. Participants ranged between the ages of 48 and 95. Participant data was collected across two studies (see Schmitter-Edgecombe et al., 2009 and Schmitter-Edgecombe et al., 2011). Interview, testing, and collateral medical information were carefully evaluated to determine whether each participant met published clinical criteria for MCI or dementia. CDR ratings were provided by certified raters.

Neuropsychological and Demographic Measures

Demographic: Age; Education; Gender
Functional Ability: Washington State University-Instrumental Activities of Daily Living (Schmitter-Edgecombe et al., 2012); Lawton IADL (Lawton & Brody, 1969)
Depression: Geriatric Depression Scale (Yesavage et al., 1983)
Mental Status: Telephone Interview For Cognitive Status (TICS; Brandt et al., 1988)
Attention/Speeded Processing: Trails A (Reitan, 1958), Symbol Digit Modalities Test (Smith, 1991)
Verbal Learning & Memory: Memory Assessment Scales (Williams, 1991); RAVLT (Lezak et al., 2004)
Visual Learning & Memory: 7/24 (Barbizet & Cany, 1968); Brief Visual Memory Test (Benedict, 1997)
Executive Functioning: Clox (Royal, et al., 1998); Trails B (Reitan, 1958); Delis-Kaplan Executive Functioning System (D-KEFS) Design Fluency (Delis et al., 2001)
Working Memory: WAIS-III Letter-Number Span and Sequencing (Wechsler, 1997)
Verbal Fluency: D-KEFS Letter and Category Fluency subtests (Delis et al., 2001)
Confrontation Naming: Boston Naming Task (Kaplan et al., 1983)
Word Knowledge: Shipley Institute of Living Scale (Zachery, 1991)

Table 1: Attributes included in models by domain

Attributes used in the machine learning analyses are listed in Table 1. Different depression, functional ability, and verbal and visual memory measures were used in the two studies for which participant data is drawn. Therefore, these scores were converted to z-scores to facilitate cross-study comparisons.

Attribute Reduction

The original datasets had 149 attributes for the CDR and diagnostic classes. Some of the variables in the original dataset were similar in nature and others have not been validated by the clinical community. As a result, there was a desire to reduce the dataset to one of more clinical

relevance. At the same time, we would like to explore whether the additional attributes strengthen the performance of the machine learning-based automated diagnostic approach. To address both of these points, we considered both the original datasets and datasets with a reduced set of attributes. The reduced-attribute datasets only include those attributes of clinical relevance that could not be subsumed by a combination of other variables. From the original datasets were reduced to 28 attributes. The reduced datasets will be referred to as Diagnosis-Reduced, CDR-Reduced, and SS-CDR-Reduced.

The impact of feature selection was also examined. In addition to reducing the dimensionality of the learning problem, feature selection can play a valuable role in determining which measures are the most critical to classification decisions. A wrapper-based feature selection approach was utilized. This approach uses an existing classifier to evaluate the feature subsets created based on their performance (accuracy) in the predictive model (Gütlein et al., 2009). Feature vector size was limited to 12 attributes, this number was determined through the initial phases in feature selection. In the initial phase, it was found that 12 attributes could increase the accuracy of the classifiers. The results reported were collected using Naïve Bayes as the base classifier. NB was chosen as it was found to have the highest degree of accuracy across all reduced datasets. The feature selection datasets will be referred to as Diagnosis-FS, CDR-FS, and SS-CDR-FS.

Models

Four machine learning algorithms were implemented in Python using Orange (Curk et al., 2005): NB, C4.5 DT, back-propagation NN, and SVM with radial basis function kernel type. Discretization was performed on the data, as not all of the classifiers were able to handle continuous variables. For the discretization step, we used entropy based discretization. In this method the discretization step determines the suitable number of intervals by recursively splitting the domain of the continuous variables to minimize the class entropy. For supervised learning, 5-fold cross validation was used for the evaluation of the models.

A self-training approach to semi-supervised learning was also utilized to increase the sample size of the CDR group by making use of the unlabeled data (i.e., when no CDR rating was assigned). Approximately 24% of the data was unlabeled. For the semi-supervised approach, each classifier determined the CDR label for a given participant. The data with exiting labels was evaluated using 5-fold cross validation. The newly labeled data was then added to the training set (80% of the originally labeled data was used for training in each fold), to help develop the model. The newly labeled data was not used in the testing phase. Rather, it was only used to help develop the models.

Model performance was assessed based on three criteria: (1) classification accuracy, (2) sensitivity, and (3) selectivity. Judging the machine learning models with the CDR datasets will allow comparison between model accuracy and the inter-rater reliability of approximately 80% found within the CDR literature (Burke et al., 1988; McCulla et al., 1989; Morris, 1997). The sensitivity and specificity for neuropsychological evaluations of cognitive impairment is between 80-90% and 44-98%, respectively (Cahn et al., 1995; Swearer et al., 1998; Wilder et al., 1995). Therefore, the performance of the models using the clinical diagnosis dataset will show clinical relevance if sensitivity and specificity are within these ranges.

Missing Attribute Values

Attribute values were missing in the datasets with a frequency of 2-4%. Missing attribute values were replaced with average values. When the class was known the missing attribute value was replaced with the average value of the given attribute for the associated class. When the class was unknown the missing attribute value was replaced with the overall average value of the attribute.

Results

Results from the feature selection process revealed that 5 of 12 selected attributes were the same for both the CDR and clinical diagnosis groups. Specifically, age, gender, education, functional ability z-score, and visual memory total correct z-score were identified as critical classification attributes for classification in both datasets. In addition to these attributes, feature selection identified total score on TICs (estimate of cognitive functioning), Trails B (executive functioning), Clox 1 (executive functioning), open dots and switching subtests of the design fluency task (executive functioning), depression z-score, and delayed verbal memory z-score as important variables for the classification of MCI, dementia, and normal older adults in the Diagnosis-Reduced dataset. Feature selection for the CDR-Reduced dataset, on the other hand identified Trails A (attention), letter, category, and switching fluency of the verbal fluency subtest (language, executive functioning), solid dots subtest of the executive functioning task (executive functioning), short-delay verbal memory z-score and long-delay visual memory as critical variables for classification of individuals with CDR = 0, 0.5, or 1.0.

The supervised learning models were tested using 5-fold cross validation. The datasets included were the Diagnosis and the CDR datasets. The semi-supervised approach was tested on the SS-CDR dataset only using a 5-fold cross validation on the originally labeled dataset. The newly labeled data was added to the training dataset, and never used in testing the model. In other words, for each fold there was an 80/20 split in originally labeled data for training and testing with the additional newly labeled data added to the training data. Please see Table 2 for results from all the datasets.

Supervised Learning: Clinical diagnosis used as class											
			Diagnosis	Diagnosis-Reduced	Diagnosis-FS				Diagnosis	Diagnosis-Reduced	Diagnosis-FS
Naïve Bayes	missing	acc	76.2%	78.5%	83.3%	SVM	missing	acc	70.7%	68.8%	60.1%
		sens	77.4%, 67%, 81.4%	79.2%, 62.9%, 87.6%	86.8%, 67%, 91.9%			sens	32.8%, 64%, 85.1%	56.2%, 75.2%, 68.8%	21.8%,62.8%, 70.7%
		spec	99.2, 80.8%, 79.3%	98.8%, 86%, 77.3%	98.8%, 91.1%, 80%			spec	100%, 78.4%, 70%	98.8%, 68.5%, 82%	97.7%, 68.1%, 66%
	not missing	acc	77.8%	77.5%	81.7%		not missing	acc	79.4%	68.8%	70.1%
		sens	88.7%, 69.1%, 79.5%	83%, 63.9%, 83.9%	88.7%, 67%, 88.2%			sens	40%, 81.5%, 90.7%	50.7%, 74.2%, 68.8%	50.6%, 73.2%, 72%
		spec	97.3%, 82.2%, 84%	96.9%, 841%, 81.3%	98.4%, 88.8%, 80.7%			spec	100%, 80.3%, 85.3%	96.7%, 67.9%, 83.3%	95.7%, 70.9%, 82.7%
Decision Tree	missing	acc	65.9%	68.5%	74.3%	NN	not missing	acc	80.3%	78.1%	82%
		sens	68.9%, 48.5%, 75.2%	77.4%, 45.4%, 79.5%	73.6%, 50.5%, 88.8%			sens	83%, 66%, 86.3%	83%, 62.9%, 85.1%	86.8%, 64.9%, 90.7%
		spec	88.4%, 80.4%, 77.3%	93.8%, 81.8%, 71.3%	95.3%, 86%, 74.7%			spec	96.5%, 86.4%, 84%	97.7%, 85%, 80%	98.1%, 90.2%, 80%
	not missing	acc	79.7%	71.1%	78.5%						
		sens	86.8%, 62.9%, 87.6%	73.6%, 46.4%, 85.1%	79.2%, 66%, 85.7%						
		spec	95.7%, 88.3%, 82%	95%, 84.1%, 71.3%	96.5%, 85.5%, 82%						

Supervised Learning: CDR score used as class											
			CDR	CDR-Reduced	CDR-FS				CDR	CDR-Reduced	CDR-FS
Naïve Bayes	missing	acc	73.9%	75%	80.1%	SVM	missing	acc	66.9%	65.1%	69.1%
		sens	85.1%, 58.1%, 64%	85.7%, 59.1%, 68%	92.2%, 61.3%, 76%			sens	71.4%, 67.7%, 0%	74.9%, 57.1%, 0%	84.1%, 54.9%, 0%
		spec	72%, 82.1%, 97.6%	71.2%, 83.2%, 98.4%	72%, 89.9%, 98.8%			spec	67.3%, 71.7%, 100%	60.1%, 74.4%, 100%	59.9%, 80.6%, 100%
	not missing	acc	81.6%	75.7%	79.4%		not missing	acc	82%	69.5%	71.7%
		sens	86.4%, 76.3%, 72%	85.7%, 59.1%, 76%	89.6%, 62.4%, 80%			sens	88.5%, 78.5%, 53%	84.6%, 48.7%, 33%	85.5%, 58.2%, 22.7%
		spec	83.9%, 84.4%, 98.8%	73.7%, 84.4%, 97.2%	72.9%, 88.3%, 98.8%			spec	83.2%, 84.3%, 99.6%	57.7%, 83.8%, 99.2%	64.2%, 81.1%, 100%
Decision Tree	missing	acc	70.2%	65.4%	74.3%	NN	not missing	acc	76.9%	75%	77.2%
		sens	80.5%, 54.8%, 64%	74.7%, 50.5%, 64%	89%, 52.7%, 64%			sens	81.8%, 65.6%, 80%	84.4%, 61.3%, 68%,	88.3%, 58.1%, 80%
		spec	72.9%, 79.3%, 95.1%	69.5%, 75.4%, 94.3%	72%, 85.5%, 95.5%			spec	74.6%, 82.7%, 98.8%	71.2%, 82.1%, 99.2%	69.5%, 87.2%, 98.8%
	not missing	acc	81.6%	68.4%	73.9%						
		sens	88.7%, 73.1%, 76%	79.2%, 49.5%, 72%	85.7%, 53.8%, 76%						
		spec	84.7%, 87.2%, 96.4%	67.8%, 79.3%, 95.5%	66.9%, 84.9%, 98%						

Semi-Supervised Learning: CDR score used as class											
			SS-CDR	SS-CDR-Reduced	SS-CDR-FS				SS-CDR	SS-CDR-Reduced	SS-CDR-FS
Naïve Bayes	missing*	acc	83.8%	77.6%	79.1%	SVM	missing*	acc	76.5%	66.2%	75.4%
		sens	83.1%, 89.2%, 68%	85%, 65.7%, 76%	90.2%, 61.3%, 76%			sens	82.1%, 72%, 0%	75.1%, 50.4%, 0%	93.6%, 51.4%, 46.7%
		spec	93.2%, 81%, 99.2%	75.4%, 83.8%, 98.8%	72%, 88.3%, 98.8%			spec	67.7%, 87.6%, 100%	51%, 85.2%, 100%	61.5%, 88.8%, 99.6%
	not missing	acc	80.1%	76.8%	76.8%		not missing	acc	86.4%	71%	74.3%
		sens	82.5%, 80.6%, 64%	85.7%, 63.5%, 72%	88.9%, 56.9%, 76%			sens	93.5%, 82.7%, 53%	80.7%, 60.2%, 6.7%	93%, 52.5%, 34.7%
		spec	87.2%, 79.9%, 98.8%	73.7%, 83.8%, 98.8%	68.6%, 87.1%, 98.8%			spec	85.6%, 88.8%, 100%	60.9%, 85.3%, 100%	61.7%, 86.5%, 100%
Decision Tree	missing*	acc	84.2%	65.1%	71.3%	NN	not missing	acc	75.7%	76.9%	75.4%
		sens	85.2%, 78.5%, 100%	74.2%, 48.4%, 71%	87.6%, 45.2%, 68%			sens	82.4%, 656%, 72%	88.3%, 62.3%, 60%	90.9%, 56.8%, 48%
		spec	84%, 89.3%, 98%	67.3%, 76.3%, 94.3%	66.9%, 85.4%, 94.7%			spec	73.7%, 81%, 99.6%	71.2%, 84.3%, 99.6%	67.7%, 84.9%, 99.2%
	not missing	acc	84.5%	68.8%	78.7%						
		sens	86.4%, 83.8%, 76%	76.8%, 53.8%, 75%	93.5%, 57%, 66%						
		spec	88.9%, 86%, 98.4%	68.2%, 78.1%, 96.8%	67.2%, 90.5%, 99.2%						

Note: missing = missing values, missing* = missing values were replace after semi-supervised learning step, not missing = missing values were replaced with average, acc = accuracy, sens = sensitivity, spec = specificity. Sensitivity and specificity have three values reported, representing the sensitivity and specificity for clinical diagnosis groups (i.e., AD, MCI, conrol) and CDR group (i.e., 0, 0.5, 1)

Table 2: Results from supervised and semi-supervised learning models for original, reduced, and feature selection data.

Discussion

The purpose of this study was to evaluate four machine learning models (i.e., NN, NB, SVM, DT), to construct classification models for distinguishing between diagnoses of cognitively healthy, MCI, and dementia and between CDR ratings of 0 (no dementia), 0.5 (very mild dementia), and 1.0 (mild dementia). Based on the results, it was observed that NB was able to obtain the highest classification accuracy regardless of class (i.e., CDR, clinical diagnosis) or supervised compared to semi-supervised learning. This finding partially fulfills our hypothesis regarding model performance. Our prediction that SVM would also perform with high rates of accuracy, on the other hand, was not supported. More specifically, the SVM model had the lowest degree of accuracy for both classes when the feature selection datasets were used. Furthermore, it was observed that as number of attributes was reduced, accuracy of SVM decreased as well. With the reduction of attributes, the SVM classifier was unable to model the data as sufficiently as it was with the complete variable set. As the number of examples between the original, reduced, and feature selection datasets stayed the same, sample size did not factor in to the decrease in accuracy.

While we did fine tune the SVM classifiers by adjusting the width parameter and the cost parameter (we explored values from 10^{-6} to 10^{6} for both parameters), there was only a slight improvement (at most 4% increase in accuracy). With the low accuracy from the SVM classifier and the high number of support vectors required, it suggests that there is not a lot of regularity in our data for each class. With a greater number of examples, which could provide a clearer view of each of the classes, the SVM classifiers should improve in accuracy. However, due to the limited number of participants in the current study, there is not enough example data for the SVM classifier.

As hypothesized, modeling for clinical diagnosis achieved better accuracies than modeling for CDR scores. This is likely due to the fact that the same attributes that were used in the machine learning models were also used to originally assign diagnostic labels. CDR scores, on the other hand, were made without overlapping influence from the attributes used in the models.

For the clinical diagnosis dataset accuracy, sensitivity, and specificity improved with the feature selected dataset. Sensitivity for MCI (64.9-67%) remained below the acceptable threshold observed in the literature (80%). This finding suggests that, while NB and NN machine learning

models achieve a high degree of accuracy, sensitivity, and specificity for diagnosing individuals with dementia or no impairment, sensitivity remains low for MCI. As individuals with MCI represent a heterogenic group of individuals, future work may be interested in determining specific attributes that differentiate between MCI and control as well as MCI and dementia.

The models for CDR classification performed with varying results. When the CDR-FS dataset was used, accuracy was 80.1% with the supervised dataset and 79.1% for the semi-supervised dataset for NB which meets the inter-rater reliability for the CDR of approximately 80%.

Examination of the sensitivity and selectivity of the models for predicting CDR revealed that the CDR = 0.5 group (proxy for MCI) had the lowest sensitivity and CDR = 0 (cognitively healthy group) had the highest. These results indicate that participants meeting CDR criteria for 0.5 represent the most difficult group for the models to accurately classify while CDR = 0 was the easiest. More participants fell within the CDR = 0 group which may partially explain why this group had the highest sensitivity and lowest selectivity results. Like MCI participants in the clinical diagnosis dataset, CDR = 0.5 represents an intermediate stage between cognitively healthy and dementia. Therefore, individuals with a CDR of 0.5 may have more variability in their performances than either individuals without cognitive impairment or those with significant cognitive impairment.

Based on the results, differences were observed between the classifier results when they included missing attribute values and when they did not have missing attribute values. In most cases, replacing the missing attributes with the average for that class improved the accuracy. By replacing the missing attribute values a more comprehensive model was achieved, with the exception of NB. The NB classifier provided stronger classification accuracy when missing values were retained. Furthermore, NB with missing values had the highest accuracy rates of all the machine learning methods regardless of dataset. Given the possibility of incomplete data in clinical practice, a method that acquires an accurate in the absence of information is extremely valuable. Furthermore, it is best to use known data rather than infer or estimate missing attribute-values when making important clinical diagnoses. Under these circumstances, NB appears advantageous relative to other machine learning models.

In about half of the cases there was a decrease in accuracy from the original datasets to the reduced datasets. The decrease in accuracy can be attributed to the reduction in attributes. For example, the original datasets contained attributes that were overlapping and experimental, whereas the reduced datasets contained only clinically validated total scores on neuropsychological tests. For example, the original datasets contained attributes that broke down a measure of global cognitive impairment into question 1, question 2 etc., whereas the reduced datasets contained only the overall score. The generality of the reduced dataset impacted the accuracy of some of the models, especially in the case of the SVM classifier. While having the original datasets created a more comprehensive model, the clinical relevance may be limited.

With an initial feature selection performed on the reduced dataset, we were able to improve the accuracy of every classifier, with the exception of the SVM classifier. In most cases, the classifiers were able to perform equivalently, or better than, the original dataset as the attributes used were only the ones determined to be critical for classification. The feature selection was done with NB as the base classifier. As the feature selection sets are optimized to the NB classifier, it makes sense that the NB classifier performs better than the other classifiers. To get a comprehensive feature selection set, other classifiers need to be used as the base classifier.

Future Work

One of the most noticeable issues with our datasets are the class imbalances. In all the datasets used, there are far more healthy participants than participants with either mild cognitive impairment, or dementia. Next we plan to explore sampling methods that could minimize the effects that the skew of the dataset are having on the results. Specifically, we are considering synthetic minority over-sampling technique (SMOTE; Chawala et al., 2002). We believe that SMOTE could improve the accuracy of by under-sampling the majority class (healthy), which will increase the sensitivity to the minority classes (MCI, dementia).

In addition to the four models explored in this study (NB, DT, NN, and SVM), another model that should be explored is an ensemble method. The ensemble method will take advantage of the multiple classifiers and should achieve a better overall predictive performance than achieved by any single classifier. It may also be of interest to explore how machine learning models compare to traditional statistical methods (e.g., logistic regression, discriminate analysis) for prediction and classification.

In regard to feature selection, the first area that should be explored involves the base classifier. In this study, we used the NB classifier as the base classifier. By only using NB as the base classifier, we may have inflated our feature selection results in favor of NB. Using the other classifiers (DT, NN, SVM) as the base classifier could provide additional insights. Another expansion involving feature selection would be to not limit the number of selected attributes to twelve. Leaving the range open may provide us with the best possible subset, independent of the number of attributes available to select. As the goal of feature selection is to find the smallest subset of attributes that maintain or improve accuracy, it leads to exploring subsets that are smaller than twelve

In addition to feature selection, other dimensionality reduction techniques should be investigated. Since feature selection provides the optimal feature subset, employing a reduction technique such as principal component analysis could provide insights as to how the neuropsychological and demographic variables work together.

Conclusions

We explored the use of neuropsychological and demographic data to predict CDR scores and clinical diagnoses through the implementation of four machine learning models (i.e., NN, NB, SVM, DT). We hypothesized that NB and SVM would provide the greatest accuracy. Based on the results, NB achieved the highest accuracy in all cases. However, our prediction that SVM would also provide a high accuracy rate, was not supported. In fact, the SVM classifier had the lowest accuracy rate when the feature selection dataset was used. We also hypothesized that because clinical diagnoses were made by including scores on neuropsychological measures used in the machine learning models the clinical diagnosis group would show classification accuracies higher than the CDR group. Based on the data, this hypothesis was supported.

We also explored the use of feature selection to reduce the number of demographic and neuropsychological data needed to make an accurate classification. We were able to determine which tests were critical when making classification for CDR scores and clinical diagnoses. Of note we were able to surpass the inter-rater reliability of 80% accuracy for classifying CDR scores and met the expected sensitivity and specificity range for each clinical diagnosis group with the exception of MCI.

The experiments reported in this paper indicate that artificial intelligence techniques can be used to automate aspects of clinical diagnosis and can provide meaningful insights into which attributes are the most valuable for this diagnosis. Continued investigation will highlight ways that these methods can be used to reduce diagnosis cost and to improve health-related decision making.

References

Barbizet, J.; and Cany, E. 1968. Clinical and psychometrical study of a patient with memory disturbances. *International Journal of Neurology* 7: 44.

Benedict, R. H. B.1997. *Brief visuospatial memory test-revised.* Odessa, FL: Psychological Assessment Resources.

Burke, W. J.; Miller, J. P.; Rubin, E. H.; Morris, J. C.; Coben, L. A.; Duchek, J.; ... and Berg, L. 1988. Reliability of the Washington University clinical dementia rating. *Archives of* Neurology 45: 31-32.

Brandt, J.; Spencer, M.; and Folstein, M. 1988. The telephone interview for cognitive status. *Cognitive and Behavioral Neurology* 2: 111-118.

Chawla, N.; Bowyer, K.; Hall, L.O.; and Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16:327-357.

Cahn, D. A.; Salmon, D. P.; Butters, N.; Wiederholt, W. C.; Corey-Bloom, J.; Edelstein, S. L.; and Barrett-Connor, E. 1995. Detection of dementia of the Alzheimer type in a population-based sample: Neuropsychological test performance. *Journal of the International Neuropsychological Society* 1: 252-260.

Chen, R.; and Herskovits, E. H. 2010. Machine-learning techniques for building a diagnostic model for very mild dementia. *Neuroimage* 52: 234-244.

Curk, T.; Demšar, J.; Xu, O; Leban, G.; Petrovič, U.; Bratko, I.; … and Zupan, B. 2005. *Microarray data mining with visual programming. Bioinformatics.* 21: 396-8.

Delis, D. C.; Kaplan, E.; and Kramer, J. H. 2001. *Delis–Kaplan executive Function System: Examiner's manual.* San Antonio, TX: The Psychological Corporation.

Kaplan, E. F.; Goodglass, H.; and Weintraub, S. 1983. *The Boston naming test.* Philadelphia: Lea & Febiger.

Lezak, M. D.; Howieson, D. B.; Loring, D. W.; Hannay, H. J.; and Fischer, J. S. 2004. *Neuropsychological Assessment (4th ed.)*. New York: Oxford University Press.

McCulla, M. M.; Coats, M.; Van Fleet, N.; Duchek, J.; Grant, E.; and Morris, JC. 1989. Reliability of clinical nurse specialists in the staging of dementia. *Archives of Neurology* 46:1210-1211.

Morris, J.C. 1997. Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *International Psychogeriatrics* 9: 173-176.

Petersen, R. C.; Doody, R.; Kurz, A.; Mohs, R. C.,; Morris, J. C.; Rabins, P. V.;… and Winblad, B.2001. Current concepts in mild cognitive impairment. *Archives of Neurology* 58: 1985-1992.

Quintana, M.; Guàrdia, J.; Sánchez-Benavides, G.; Aguilar, M.; Molinuevo, J. L.; Robles, A.; ... and for the Neuronorma Study Team. 2012. Using artificial neural networks in clinical neuropsychology: High performance in mild cognitive impairment and Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology* 34: 195-208.

Reitan, R. M. 1958. Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills* 8: 271–276.

Royall, D. R.; Cordes, J. A.; and Polk, M. 1998. CLOX: An executive clock drawing task. *Journal of Neurology, Neurosurgery & Psychiatry* 64: 588-594.

Shankle, W. R.; Mani, S.; Dick, M. B.; and Pazzani, M. J. 1998. Simple models for estimating dementia severity using machine learning. *Studies in Health Technology and Informatic*s 16: 472-476.

Schmitter-Edgecombe, M.; McAlister, C.; and Weakley, A. 2012. Naturalistic assessment of everyday functioning in individuals with mild cognitive impairment: the day out task. *Neuropsychology* 26: 631-641.

Schmitter-Edgecombe, M.; Parsey, C.; and Cook, D. 2011. Cognitive correlates of functional performance in older adults: Comparison of self-report, direct observation and performance-based measures. *Journal of the International Neuropsychological Society* 17: 853-864.

Schmitter-Edgecombe, M.; Woo, E.; & Greeley, D. 2009. Characterizing multiple memory deficits and their relation to everyday functioning in individuals with mild cognitive impairment. *Neuropsychology* 23: 168-177.

Swearer, J. M.; O'Donnell, B. F.; Kane, K. J.; Hoople, N. E.; and Lavoie, M. (1998). Delayed recall in dementia: sensitivity and specificity in patients with higher than average general intellectual abilities. *Cognitive and Behavioral Neurology* 11: 200-206.

Wilder, D.; Cross, P.; Chen, J.; Gurland, B.; Lantigua, R. A.; Teresi, J.; ... and Encarnacion, P. 1995. Operating characteristics of brief screens for dementia in a multicultural population. *The American Journal of Geriatric Psychiatry* 3: 96-107.

Smith, A. 1991. *Symbol digit modalities test.* Los Angeles: Western Psychological Services.

Williams, J. M. 1991. *Memory assessment scales*. Odessa, FL: Psychological Assessment Resources.

Yesavage, J. A.; Brink, T. L.; Rose, T. L.; Lum, O.; Huang, V.; Adey, M.; and Leirer, V. O. 1983. Development and validation of a geriatric depression screening scale: A preliminary report. *Journal of Psychiatric Research* 17: 37-49.

Zachary, R. A. 1991. *Shipley Institute of Living Scale—Revised manual.* Los Angeles: Western Psychological Services.