Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia

Jennifer Williams, Alyssa Weakley
Washington State University, Pullman, WA
jen_williams@wsu.edu, alyweakley@gmail.com

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, 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 explored. Our results show that the NB classifier provided the best accuracies, while the SVM classifier proved to provide some of the lowest accuracies. We also illustrate that with the use of feature selection, accuracies can be improved. We conclude that 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, which may have significant implications for the future of health care.

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, we explore the use of machine learning algorithms to automate the analysis of medical data that is used to perform assessment of cognitive health. A few prior studies have been successful in using machine learning algorithms to accurately classify participants with cognitive impairment. For example, 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 30 minutes and requires interviews with both the individual with cognitive impairment and a significant other, making it impractical in some clinical settings. While the authors were unable to exceed the CDR's inter-rater reliability of approximately 80% (McCulla et al., 1989; Burke et al., 1998), they concluded that the reduction in accuracy did not outweigh the decreased information needed to obtain a CDR score or time required to classify individuals. 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 the dementia 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), defined as the transitional state between normal aging and dementia (Petersen et al. 2001), Alzheimer's disease (AD), 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 participants, the model classified all individuals correctly. While the later 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 AD 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 AD [CDR = 0.5; proxy for MCI], and AD [CDR = 1]) and (2) clinical diagnoses (i.e., cognitively healthy, MCI, AD) through the implementation of four well-known machine learning models: NN, SVM, naïve Bayes (NB), and decision tree.

In addition to classifying individuals, a secondary goal of the project is 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 dataset used clinical diagnosis as the class that was being predicted. This dataset will be referred to as the Diagnosis dataset. The second dataset used CDR scores as the class that was predicted, 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 final dataset will be referred to as the SS-CDR dataset.

Participants

Participants in Diagnosis dataset were placed into one of three target class values: AD, MCI, or cognitively healthy older adult. The participants included 55 individuals with AD (24 female, 31 male), 101 individuals with MCI (57 female, 44 male), and 179 cognitively healthy older adult control participants (126 female, 53 male). Participants in the CDR datasets included 174 individuals (118 females, 56 males) that received a CDR of 0.0 (no dementia), 100 individuals (60 females, 40 males) with a CDR of 0.5 (very mild dementia), and 27 individuals (9 females, 18 males) 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. 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 AD. CDR ratings were provided by certified raters.

Measures

Attributes used in the machine learning analyses consisted of demographic variables (age, education, and gender), functional ability, depression (Geriatric Depression Scale [GDS]; Yesavage et al., 1983), mental status (Telephone Interview for Cognitive Status [TICS]; Brandt et al., 1988), and scores on neuropsychological tests. Performances on the following neuropsychological tests were included: attention and speeded processing (Symbol Digit Modalities Test [SDMT]-Oral and Written subtests; Smith, 1991; Trail Making Test, Part A; Reitan, 1958),verbal learning and delayed memory (RAVLT; Lezak et al. 2004; or the Memory Assessment Scales [MAS]; Williams 1991, depending on the study sample), visual learning and memory (7/24; Barbizet and Cany 1968; or the Brief Visual Memory Test [BVMT]; Benedict 1997, depending on the study sample), executive functioning (Trail Making Test, Part B; Reitan 1958, Clox 1; Royall et al., 1998; Design Fluency subtest of Delis-Kaplan Executive Functioning System [D-KEFS]; Delis et al., 2001), working memory (WAIS-III Letter-Number span and sequencing; Wechsler 1997), verbal fluency (verbal fluency subtest from the D-KEFS; Delis et al., 2001), confrontation naming (BNT; Kaplan et al., 1983), and word knowledge (Shipley Institute of Living Scale; Zachary, 1991). Different GDS, 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 and 156 attributes for the CDR and diagnostic classes, respectively. 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 attributes were reduced to 28 and 29, respectively. The extra attribute in the Diagnostic-Reduced dataset is the CDR score, which is the target attribute in the CDR dataset. The reduced datasets will be

referred to as Diagnosis-Reduced, CDR-Reduced, SS-CDR-Reduced,

The impact of a 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. The experimental results reported were collected based on preliminary feature selection using Naïve Bayes as the base classifier. The feature selection datasets will be referred to as Diagnosis-FS, CDR-FS, SS-CDR-FS.

Models

Four machine learning algorithms were implemented in Python using Orange (Curk et al., 2005): NB, C4.5 decision tree, back-propagation NN, and SVM. Discretization was performed on the data, as not all of the classifiers were able to handle continuous variables.

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., 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 newly labeled data was then added into the training set. The training set was then evaluated on the test data to determine if classification accuracy improved.

Performance was assessed based on the criteria: (1) classification accuracy, (2) sensitivity, and (3) selectivity.

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 7 of 12 selected attributes were the same for both the CDR and clinical diagnosis groups. Specifically, education, TICS (global cognitive status), trails B (executive functioning) design fluency (executive functioning), functional ability z-score, depression z-score, and short delay verbal memory z-score were identified as critical attributes for classification in both datasets. In addition to these attributes, feature selection identified total CDR score, age, letter-number span (attention, working memory), and Clox1 and 2 (executive functioning, constructional abilities) as important variables for classification of MCI, AD, and normal older adults in the Diagnostic-Reduced dataset. Feature selection for the CDR-Reduced dataset, on the other hand, identified the Shipley Institute of Living Scale (fund of knowledge), letter fluency of the verbal fluency subtest (language, executive functioning), BNT (language), and visual and verbal delayed memory z-scores as critical variables for classification of individuals as of 0, 0.5 or 1.

Classification accuracies among the three different types of datasets (i.e., original, reduced, feature selection) varied (see Table 1). Using the supervised approach for clinical diagnosis, an improvement in classification accuracy was observed as the attribute list was reduced. More specifically, accuracy improved or remained unchanged when the Diagnosis-Reduced dataset was employed (76.2-90.4%) compared to the Diagnosis dataset (74.3-89.7%), and improved further with the Diagnosis-FS dataset (81.3-93.3%) for each of the models with the exception of SVM. With SVM a reduction in attributes was observed from each dataset to the next.

Using the supervised approach to CDR classification improvements in accuracies were observed from the CDR dataset (73.9-80.9%) to the CDR-Reduced dataset (69.2-75.7%) for NB without missing values, decision tree with missing values and NN. Accuracy improvement was also observed between the CDR-Reduced dataset (69.2-75.7%) and the CDR-FS dataset (67.3-78.7%) for each of the models with the exception of SVM without missing variables (65.4-75.4). Using the semi-supervised approach to CDR classification improvements in accuracies were observed from the SS-CDR dataset to the SS-CDR-Reduced dataset for NB with missing values, decision tree with missing values and SVM with missing values. A decrease in performance was observed between the SS-CDR and SS-CDR-Reduced datasets for NB decision tree, SVM without missing values, and NN. Accuracy improvement was realized between the SS-CDR-Reduced dataset (72.2-81.4%) and SS-CDR-FS dataset (73.0-83.4%) for each of the models with the exception of SVM.

Due to space constraints following are results from the feature selection approach. Please see Table 1 for comparisons amongst the other datasets. For the Diagnostic-FS dataset, NB proved to have the highest classification rate with 92% accuracy when missing values were present and 93.3% when missing values were replaced with class averages. Sensitivity and specificity were high, ranging from 87.6 to 95.7% and 92.7 to 99.6%, respectively. C4.5 decision tree performed with 81.3% classification accuracy with missing values and 91% without missing values. NN performed with 91.6% accuracy without missing values. Of the machine learning methods, SVM provided the lowest classification rate at 65.6% accuracy with missing values and 79.4% without. SVM sensitivity ranged from as low as 15.1 for AD classification to 89.7% for MCI.

Supervised Learning: Clinical diagnosis used as class											
			Diagnosis	Diagnosis-Reduced	Diagnosis-FS				Diagnosis	Diagnosis-Reduced	Diagnosis-FS
Naïve Bayes	missing	acc	80.7%	83.9%	92%	SVM	missing	acc	79.1%	66.2%	65.6%
		sens	77.4%, 79.4%, 82.6%	84.9%, 74.2% 89.4%	88.7%, 87.6%, 95.7%			sens	37.7%, 81.4%, 91.3%	18.9%, 78.4%, 74.5%	15.1%, 79.4%, 73.9%
		spec	81.8%, 88.7%, 98.4%	84.7%, 88.8%, 98.8%	99.2%, 94.4%, 92.7%			spec	99.6%, 80.8%, 84.7%	90.7%, 81.3%, 72.7%	87.6%, 86%, 70%
	not missing	acc	85.2%	85.2%	93.3%		not missing	acc	86.1%	78.8%	79.4%
		sens	82.5%, 85.1%, 90.6%	77.3%, 88.7%, 88.8%	88.7%, 90.7%, 96.3%			sens	49.1%, 91.8%, 95%	60.4%, 89.7%, 78.3%	62.3%, 89.7%, 78.9%
		spec	86.9%, 90.7%, 96.9%	89.3%, 90%, 96.9%	99.6%, 94.9%, 94%			spec	100%, 84.6%, 93.3%	92.6%, 82.7%, 93.3%	98.4%, 81.3%, 86.7%
Decision Tree	missing	acc	74.3%	76.2%	81.3%	NN	not missing	acc	89.7%	90.4%	91.6%
		sens	66%, 53.6%, 89.4%	69.8%, 55.7%, 90.7%	71.7%, 63.9%, 95%			sens	83%, 81.4%, 96.9%	83%, 84.5%, 96.3%	81.1%, 86.6%, 98.1%
		spec	89.9%, 87.9%, 81.3%	91.5%, 90.7%, 78.7%	96.5%, 92.1%, 78.7%			spec	97.3%, 93.9%, 92%	98.1%, 93.5%, 92.7%	98.8%, 94.4%, 92.7%
	not missing	acc	88.7%	90.1%	91%						
		sens	81.1%, 82.5%, 95%	79.2%, 87.6%, 95%	79.2%, 86.6%, 97.5%						
		spec	98.4%, 92.1%, 90.7%	98.8%, 91.6%, 93.3%	99.6%, 94.9%, 92.7%						

Supervised Learning: CDR score used as class											
			CDR	CDR-Reduced	CDR-FS				CDR	CDR-Reduced	CDR-FS
Naïve Bayes	missing	acc	73.9%	75%	76.8%	SVM	missing	acc	64.3%	64.3%	67.3%
		sens	85.1%, 58.1%, 64%	85.7%, 59.1%, 68%	91.6%, 55.9%, 64%			sens	85.1%, 43%, 16%	87%, 43%, 4%	89%, 44.1%, 20%
		spec	72%, 82.1%, 97.6%	71.2%, 83.2%, 98.4%	69.5%, 87.7%, 98%			spec	46.6%, 81%, 10%	45.8%, 81.6%, 100%	49.2%, 83.8%, 100%
	not missing	acc	81.6%	75.7%	78.7%		not missing	acc	80.9%	69.2%	66.2%
		sens	86.4%, 76.3%, 72%	85.7%, 59.1%, 76%	88.3%, 62.4%, 80%			sens	89.6%, 73.1%, 56%	88.3%, 47.3%, 32%	89%, 39.8%, 24%
		spec	83.9%, 84.4%, 98.8%	73.7%, 84.4%, 97.2%	73.7%, 87.2%, 98.4%			spec	78%, 86%, 99.6%	55.1%, 83.8%, 99.2%	42.4%, 86.6%, 100%
Decision Tree	missing	acc	70.2%	65.4%	72.4%	NN	not missing	acc	77.2%	75%	76.1%
		sens	80.5%, 54.8%, 64%	74.7%, 50.5%, 64%	83.1%, 55.9%, 68%			sens	81.8%, 67.9%, 84%	85.1%, 61.3%, 68%	87.7%, 57%, 76%
		spec	72.9%, 79.3%, 95.1%	69.5%, 75.4%, 94.3%	68.6%, 82.1%, 97.6%			spec	76.3%, 82.2%, 99.2%	71.2%, 82.1%, 99.2%	69.5%, 86.6%, 98%
	not missing	acc	80.9%	69.2%	73.9%						
		sens	89.6%, 73.1%, 56%	88.3%, 47.3%, 32%	85.7%, 53.8%, 76%						
		spec	78%, 86%, 99.6%	55.1%, 83.8%, 99.2%	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	78.7%	81.4%	83.1%	SVM	missing	acc	69.7%	74.2%	69.9%
		sens	88.2%, 67.5%, 65.8%	90.5%, 67.5%, 76.3%	94.1%, 66.7%, 75.7%			sens	89.5%, 51.3%, 0%	91.5%, 59.7%, 0%	88%, 49.6%, 28.6%
		spec	73.3%, 86.7%, 99.4%	76.8%, 88.7%, 99.1%	76%, 91.2%, 99.4%			spec	45.9%, 87.9%, 100%	53.1%, 89.9%, 100%	51.4%, 85.6%, 100%
	not missing	acc	88.2%	80.3%	83.4%		not missing	acc	86%	74.2%	69.9%
		sens	89.7%, 91.9%, 68.4%	86.6%, 69.2%, 81.6%	90.6%, 70.1%, 86.5%			sens	96.2%, 87.2%, 6.9%	90%, 62.2%, 0%	90.4%, 53.1%, 2.9%
		spec	87.6%, 90.6%, 100%	77.4%, 86.6%, 99.1%	79.9%, 90%, 98.7%			spec	76.7%, 93.3%, 100%	54.5, 89%, 100%	43.9%, 90.1%, 100%
Decision Tree	missing	acc	68.3%	72.5%	78.9%	NN	not missing	acc	77.8%	73.9%	75.5%
		sens	79.4%, 52.1%, 62.8%	78.8%, 60.5%, 75%	90.6%, 62.2%, 68.6%			sens	86.1%, 60.7%, 77.8%	86%, 47.2%, 71.4%,	84.1%, 59.3%, 77.8%
		spec	64.2%, 81%, 96.8%	74.1%, 80.2%, 97.1%	75.3%, 88.2%, 97.2%			spec	73%, 88.9%, 99.8%	61.9%, 86.4%, 97.9%	71.1%, 85.6%, 96.9%
	not missing	acc	81.2%	72.2%	73%						
		sens	86.1%, 72.3%, 83.7%	77.8%, 63.2%, 70.5%	85.6%, 52.1%, 71.4%						
		spec	84%, 86.9%, 96.8%	74.7%, 78.5%, 97.8%	70.1%, 84%, 96.3%						

Note: missing = missing values, 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 1: Results from supervised and semi-supervised learning models for original, reduced, and feature selection data.

Similar to the Diagnosis-FS dataset, when the CDR-FS dataset was explored NB proved to have the highest accuracy rate. Specifically, with missing values the classification accuracy was 76.8% and 78.7% without, which is marginally below the CDR inter-rater reliability of 80%. C4.5 decision tree's classification rate was 72.4% with missing values and 73.9% without. NN performed with 76.1% accuracy without missing values. Of the models, SVM had the lowest classification accuracy at 67.3% with missing values and 66.2% without.

Using the SS-CDR-FS dataset NB, once again, proved to be the best classifier with accuracies of 83.1% with missing data, and 83.4% without. Of note, these accuracies surpass the CDR inter-rater reliability of 80%. C4.5 performed with 78.9% accuracy with missing data and 73.0% without. NN performed with 75.5% accuracy without missing values. Of the classifiers, SVM had the lowest classification accuracy at 69.9% with missing data, and 69.9% without. Of note, specificity for AD was 100% using the SVM model.

Discussion

The purpose of this study was to evaluate four machine learning models (i.e., NN, NB, SVM, decision tree), to construct classification models for distinguishing between diagnoses of cognitively healthy, MCI, and AD and between CDR ratings of 0 (no dementia), 0.5 (very mild dementia), and 1 (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. Of note, 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.

The accuracy, sensitivity, and selectivity of the machine learning models with clinical diagnosis as the class are considered to be robust. These results suggest that machine learning models are successful at accurately classifying individuals as either having no cognitive impairment, having mild cognitive difficulties, or falling within the domain

of dementia. It should be noted, however, that a validation problem exists within this model for classification. Specifically, the attributes used in the machine learning models were also used in the original diagnosis of participants. Since the models were then asked to classify individuals based upon this diagnosis a problem of validating the original classification emerges based on the variables used in both cases.

To circumvent this issue, classification of CDR scores were also explored. The models for CDR classification performed with varying results. When the CDR-FS dataset was used, accuracy fell within the range of 66.2% to 78.7% which is just below the inter-rater reliability for the CDR of approximately 80% (McCullen, 1988; Burke et al., 1998). As the data for feature selection is preliminary we are confident with a full feature selection process that the machine learning models will be able to surpass the inter-rater reliability. When the SS-CDR-FS dataset was used, accuracy surpassed this inter-rater reliability for the NB and NN models.

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, while selectivity was highest for CDR = 1 (AD) and lowest for CDR = 0 (cognitively healthy). 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. Since CDR = 0.5 represents an intermediate stage between cognitively healthy and dementia, 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. It is interesting to note, however, that the same pattern was not observed for the MCI group when clinical diagnosis was used for class. This may suggest that feature selected variables used to classify individuals in this intermediate range of cognitive impairment are more sensitive for the diagnosis of MCI then classification of CDR = 0.5.

As hypothesized, modeling for clinical diagnosis achieved better accuracies than modeling for CDR scores. The Diagnostic dataset also improved in accuracy with each dimensionality reduction step (with the exception of the SVM classifier). The dataset attributes only differed in CDR score between classes (i.e., CDR, clinical diagnosis). Therefore, the improved accuracy observed in clinical diagnosis as class over CDR may suggest the importance of the CDR score when predicting clinical diagnosis. As aforementioned, the original diagnoses used were derived from the attributes used in the model which also likely improved the accuracy observed in the clinical diagnosis as compared to CDR.

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.

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 TICS into TICS1, TICS2, etc., whereas the reduced datasets contained only the total TICS 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. There was the greatest increase in accuracy with the supervised learning model for the clinical diagnosis class.

Future Work

In this study, there were four models explored: NB, decision tree, 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

There are two areas in future work that involve feature selection. The first area that should be explored involves the base classifier. In this study, we used the naïve Bayes classifier as the base classifier. Using the other classifiers implemented in this study (decision tree, 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.

Another area related to feature reduction, would be to test other dimensionality reduction techniques. While our feature selection method provides the optimal feature subset, employing another reduction technique such as principal component analysis could provide other insights as to how 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, decision tree). 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. Through preliminary results, we were able to determine which tests were critical when making classification for CDR scores and clinical diagnoses. As the results from the feature selection is preliminary we are confident with a full feature selection will be able to surpass the inter-rater reliability, which is 80% accuracy for classifying CDR scores. With regard to classifying clinical diagnoses, our feature selection results were able to achieve accuracies that surpassed 80%, with the exception of the SVM classifier.

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.

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 46:1210–1, 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.; 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.

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.