Detecting Very Early Stages of Dementia from Normal Aging with Machine Learning Methods

William Rodman Shankle[1], Subramani Mani[2], Michael J. Pazzani[2] and Padhraic Smyth[2]

[1] Departments of Neurology and Information and Computer Science, University of California at Irvine, Irvine CA 92697-5100 *rshankle@uci.edu*

[2] Department of Information and Computer Science, University of California at Irvine, Irvine CA 92697-3425 *(mani,pazzani,smyth)@ics.uci.edu*

Abstract. We used Machine Learning (**ML**) methods to learn the best decision rules to distinguish normal brain aging from the earliest stages of dementia using subsamples of 198 normal and 244 cognitively impaired or very mildly demented (Clinical Dementia Rating Scale=0.5) persons. Subjects were represented by their age, education and gender, plus their responses on the Functional Activities Questionnaire (FAQ), the Mini-Mental Status Exam (MMSE), and the Ishihara Color Plate (ICP) tasks. The ML algorithms applied to these data contained within the electronic patient records of a medical relational database, learned rule sets that were as good as or better than any rules derived from either the literature or from domain specific knowledge provided by expert clinicians. All ML algorithms for all runs found that a single question from the FAQ, *the forgetting rule*, ("Do you require assistance remembering appointments, family occasions, holidays, or taking medications?") was the only attribute included in all rule sets. CART's tree simplification procedure always found that just the forgetting rule gave the best pruned decision tree rule set with classification accuracy (93% sensitivity and 80% specificity) as high as or better than any other decision tree ruleset. Comparison with published classification accuracies for the FAQ and MMSE revealed that including some of the additional attributes in these tests actually worsen classification accuracy. Stepwise logistic regression using the FAQ attributes to classify dementia status confirmed that *the forgetting rule* gave a much larger odds ratio than any other attribute and was the only attribute included in all of the stepwise logistic regressions performed on 33 random samples of the data. Stepwise logistic regression using the MMSE attributes identified two attributes which occurred in all 33 runs and had by far the highest odds ratio. In summary, ML methods have discovered that the simplest and most sensitive screening test for the earliest clinical stages of dementia consists of a single question, the forgetting rule.

1 Introduction

In this paper, we apply ML methods to the detection of the earliest stages of dementia due to Alzheimer's disease and other causes. Machine learning (**ML**) can generate classification rules where the data include the known classification of each case. The application of ML methods in the domain of medicine has been relatively infrequent because of difficulty in accessing medical data electronically. Artificial intelligence approaches to medicine started with knowledge-based systems, which learn from human experts, not data. Beginning with the expert systems of the seventies (MYCIN [28], PUFFS), followed by Bayesian systems of the late eighties and early nineties (*ACORN* [12],*PATHFINDER* [5]), these knowledge-based systems generated much enthusiasm and hope. But there are very few such actual systems in routine clinical use. Another approach starting in the mid eighties, sought to make use of real data and a domain model for knowledge acquisition and rule learning[3],[18], [15]. *KARDIO*[20] is an expert system for evaluation of electrocardiograms based on this approach. With increasing availability of electronic medical records, machine learning has the potential to become a valuable adjunct to clinical decision-making. There has been some recent effort in this direction[2].

Dementia is defined as multiple cognitive impairments with loss of related functional skills without altered consciousness. Most demented patients do not see a physician for the problem of memory loss until four years after symptom onset [7], which usually relates to the patient's social embarrassment about having a memory problem. Additionally, community physicians commonly do not detect dementia [10] or misidentify it [21] in its earliest stages when patients are seeing them for other reasons. At the mid stages of the disease, physicians are less able to slow the progression and minimize debilitating behavioral effects of the dementia. As an example of an intervention which might have greater value if started earlier in the disease, Lubeck et al. [16] reported a 17% reduction in the $200,000 cost of AD patient care using central cholinergic agonists (Tacrine). A simple, unobtrusive method for detecting dementia early in the disease's course would help get patients to seek early evaluation and treatment, resulting most probably in preserved quality of life and reduced financial burden to family and health care providers. The Agency for Health Care Policy Research (**AHCPR**) clinical practice guidelines for the assessment and recognition of Alzheimer's disease and related disorders [30] recommends two simple tests, the Functional Activities Questionnaire (**FAQ** [24]), and the six-item Blessed Orientation, Memory and Concentration test (**BOMC** [8]), to screen for dementia after excluding delirium and depression. We recently reported that the use of Machine Learning (**ML**) methods in conjunction with the FAQ and the BOMC markedly improved sensitivity in detecting dementia in a sample of 609 normal, cognitively impaired, and demented subjects when compared with published scoring criteria [27].

In this paper, we focus on discriminating the effects of normal aging on cognition from the very early stages of dementia because early detection is potentially very important for improving quality of life, and reducing total health care costs to family and society. To do this, we used the AHCPR-recommended screening instrument, the FAQ, plus the Folstein Mini-Mental Status Exam [9] and two items from the BOMC (**MMSEPLUS**) and Ishihara Color Plates (**ICP** [11]) in conjunction with several ML methods, and compared these results to those using published scoring criteria for the same set of data from the same set of subjects. Other items of the BOMC did not need to be considered in addition to the MMSE since the rest of the BOMC is a subset of the MMSE.

2 Methods

2.1 Sample Description

The total sample consisted of the initial visits of 198 cognitively normal and 244 cognitively impaired or very mildly demented (Clinical Dementia Rating Stage ≤ 0.5) subjects seen at the University of California, Irvine Alzheimer's Disease Research Center (ADRC). Patients received a complete diagnostic evaluation consisting of patient and caregiver interviews, general physical and neurological exam, two hours of cognitive testing including the CERAD [29] neuropsychological battery and other selected tests, routine laboratory testing for memory loss, and magnetic resonance neuroimaging with or without single photon emission with computed tomography. Control subjects were either community volunteers or unaffected spouses of patients, and received an abbreviated, 45 minute version of the patient cognitive battery, which consisted of the CERAD plus measures of activities of daily living. They did not receive a medical exam, laboratory testing or neuroimaging unless cognitive or functional testing suggested an impairment. The number of subjects available for the various analyses varied because of missing data. We also performed logistic regressions of the MMSEPLUS and FAQ attributes. The sample sizes for each screening test appears in Table 1.

Table 1. Characteristics of the UCI ADRC Sample of this study

Attribute	Normal			Impaired			Total		
	N	Mean	Std.	N	Mean	Std.	N	Mean	Std
Age*	196	67.2	11.8	278	68.2	10.9	474	67.6	11.3
% Female*	198	71	46	274	43	50	472	59	49
Education(yrs)	140	15.0	2.7	274	15.3	3.2	414	15.2	3.0
FAQ	137	0.2	0.8	211	7.6	6.2	348	5.1	6.1
MMSEPLUS	198	29.2	0.9	227	24.8	5.5	425	26.6	4.8
ICP	133	13.7	1.7	179	11.1	4.1	312	11.9	3.7

* T-test unpaired sample with unequal variance was significant at $P < 0.001$

Classification of Dementia Status

The diagnosis of dementia status, using DSM-IV criteria [1], was based on a review of all the data by the neurologist and neuropsychologist during their diagnostic review session. Each subject was categorized as either unimpaired, cognitively impaired but not meeting criteria for dementia, or demented. A classification of *dementia* required the presence of multiple cognitive impairments plus functional impairments resulting from the cognitive impairments in the absence of delirium or other non-organic etiologies such as major depression. They were also classified by dementia severity using standard criteria for the Clinical Dementia Rating Scale (**CDRS** [19]), in which 0 = normal, 0.5 = questionably or very mildly demented, and 1-5 indicate increasing severity of dementia. Control subjects showing cognitive impairment or very mild dementia (CDRS ≤ 0.5) were included in the cognitively impaired/very mildly demented sample, which we will refer to as the *impaired* group from here on; patients who tested normally were included in the cognitively normal sample; subjects with delirium were excluded from the analysis. Table 1 shows the sample characteristics.

FAQ, MMSEPLUS, and ICP tests

The FAQ (total score ranges from 0 (normal) to 30 (severely disabled)) consists of ten questions about basic and more complex activities of daily life. The answers to these questions were extracted from the UCI ADRC relational database of over 1,200 variables per subject-visit to compute the FAQ total and item scores. The AHCPR recommends using total FAQ scores of 9 or higher for detecting impairment. Pfeffer[24] found a total FAQ score of 5 or higher to be most sensitive as a second stage screen in discriminating normal vs. questionably demented subjects. We examined the sensitivity and specificity of total FAQ scores from 1 to 30 without ML methods. With ML methods, we used age, sex, education, and all FAQ attributes with and without the FAQ total score. The description for how these runs were performed is in the Machine Learning Methods section.

The MMSEPLUS consists of 19 questions from the MMSE regarding orientation for time and place, registration, attention, short-term recall, language, and drawing, plus two attributes from the BOMC test (recall of an address and number of trials to correctly repeat the address twice), which were added because of potential sensitivity in detecting early dementia. The MMSE ranges from 0 (severely impaired) to 30 (no impairment). The occurrence of dementia increases with advancing age and decreases with increasing educational level. Depending upon a subject's age and education, a total MMSE score of 24 or higher is used to classify a subject as normal [22, 4]. We examined the sensitivity and specificity of total MMSE scores from 1 to 30 without ML methods. We then aggregated the individual MMSE attributes reflecting short-term recall, orientation to time, and orientation to place into three aggregate attributes respectively. The MMSEPLUS attributes therefore consisted of the three MMSE aggregate attributes, individual MMSE attributes reflecting registration, attention and drawing, and the two BOMC attributes. These attributes plus age, sex and education were used with ML methods to classify normal and impaired subject samples.

The ICP consists of 21 pseudoisochromatic plates, 15 with noisy numbers and 6 with noisy trails embedded in a noisy background. The subject reads the number or traces the trail and is scored by the examiner as correct (1) or incorrect (0). The instruction for the number-naming task is, "If you see a number on the plate, tell me what it is", and that for the trail-tracing task is, "If you see the trail, trace it from beginning to end." Because it is such a simple task and it appears to discriminate among Alzheimer's, Vascular dementia and Normal aging subjects [17], we included it as a potential screening test. However, examination of the ML classification results with the 21 ICP attributes (see Table 2), showed that it is not sufficiently sensitive for detecting very mild dementia. Therefore, the ICP was removed from further consideration as a candidate for screening.

2.2 Machine Learning Methods

Specific algorithms We concentrated on decision tree learners, rule learners and the Naive Bayesian classifier. Decision trees and rules generate clear descriptions of how the ML method arrives at a particular classification. The Naive Bayesian classifier was included for comparison purposes. MLC++(Machine Learning in C++) is a software package developed at Stanford University [26] which implements commonly used machine learning algorithms. It also provides standardized methods of running experiments using these algorithms. C4.5 is a decision tree generator and C4.5rules produce rules of the form, *if..then* from the decision tree [25]. Naive Bayes is a classifier based on Bayes Rule. Even though it makes the assumption that the attributes are condtionally independent of each other given the class, it is a robust classifier and serves as a good comparison in terms of accuracy for evaluating other algorithms [6]. FOCL [23] is a concept learner which can incorporate a user provided knowledge of two types. First, when provided with a guideline or protocol directly, FOCL has the capacity for revision if the guidelines produce better classification rules than that produced from exploration of the data. Second, FOCL can accept information on each nominal variable indicating which values of the variable increase the probability of belonging to a class (such as impaired) and information on each continuous variable on whether higher or lower values of the variable increases the probability of belonging to a class. We call this, "constrained FOCL", in the experimental results. FOCL can also learn from the data only, without an initial input of constraints or guidelines. We call this, "unconstrained FOCL", in the experimental results. CART [13] is a classifier which uses a conservative tree-growing algorithm that minimizes the standard error of the classification accuracy based on a particular tree-growing method applied to a series of training subsamples. We ran CART 10 times on randomly selected 2/3 training sets and 1/3 testing sets. For each training set, CART built a classification tree where the size of the tree was chosen based on cross-validation accuracy on this training set. The test accuracy of the chosen tree was then evaluated on the unseen test set.

2.3 Treatment of missing data

We used each ML's method for handling missing data. In C4.5 missing attributes are assigned to both branches of the decision node, and the average of the classification accuracy is used for these cases. In the Naive Bayesian Classifier, missing values are ignored in the estimation of probabilities. In FOCL, any test on a missing value is treated as false. Therefore, it attempts to learn a set of rules that tolerates missing values in some variables. CART uses surrogate tests for missing values.

2.4 Training and Testing Samples

We ran experiments in which data from the FAQ, MMSEPLUS, and ICP tests were used separately by each learning algorithm. The samples for the FAQ, MMSEPLUS, and ICP ML analyses mostly overlapped but the sizes differed due to different patterns of missing data. For the FAQ there were 348 instances—137 cognitively normal and 211 impaired; for the MMSEPLUS there were 425 instances—198 normal and 227 impaired; for the ICP there were 312 instances—133 normal and 179 impaired. We cross-validated the results in the following manner. The complete sample of each screening test was used to generate 20 non-overlapping training and testing sets in a 2/3 to 1/3 ratio, with random sampling of the training set. The algorithms were trained on the training set and the resulting decision tree then classified the unseen testing set. The classification accuracy is hence the mean of the accuracies obtained for the twenty runs of the testing set. An example of one decision tree rule-set appears in figure 1.

Rule 1: age > 56 *and* job $> 2 \Rightarrow$ class **impaired**
Rule 2: money > 0 *and* forget $> 0 \Rightarrow$ class **impaired**
Rule 3: gender $= 0$ *and* age > 56 *and* forget $> 0 \Rightarrow$ class **impaired**
Rule 4: age > 56 *and* age ≤ 64 *and* forget $> 0 \Rightarrow$ class **impaired**
Rule 5: age > 73 *and* forget $> 0 \Rightarrow$ class **impaired**
Rule 6: forget $\leq 0 \Rightarrow$ class **normal**
Rule 7: Default $\Rightarrow$ class **impaired**

Fig. 1. A C45rule Set

Nonsense Rules It is possible for ML methods to generate a rule which makes no domain sense (**nonsense rule**). The rule sets generated by the various ML methods were inspected for their clinical sense by an ADRC staff neurologist. After identifying the nonsense rules, we used FOCL to incorporate domain-specific knowledge that would prevent (constrain) such rules from occurring. We then compared classification performance of the constrained vs. unconstrained runs using FOCL to see how performance was affected. An example of a decision tree with a nonsense component follows:

```
forget > 0 (having trouble):
|   age <= 52 :
|   |   edulevel > 16 : normal (4.0)
|   |   edulevel <= 16 :
|   |   |   SHOP <= 0 (no trouble shopping): impaired (5.6)
|   |   |   SHOP > 0 (having trouble shopping): normal (2.0)
```

In this example, eight persons (5.6+2.0) were forgetful, 52 years old or younger, and had 16 or fewer years of education. Among them, those who could shop were classified as impaired while those who required assistance to shop were classified as normal: this is a *nonsense rule*, which arises because of insufficient examples covering the circumstances specified by the nonsense rule. As becomes apparent later, the appearance of such nonsense rules should encourage one to gather more data, to constrain the ML method with domain-specific knowledge, or to search for a reduced rule-set using pruning techniques.

2.5 Logistic Regression Methods

50% random samples of each class were selected 33 times, and analyzed with stata's stepwise logistic regression, which estimates the odds ratios that independently contribute to the model for each run. The FAQ and MMSE were separately regressed against dementia status, and the attributes with the largest odds ratios in each run were identified.

3 Results

We examined the sensitivity (probability of correctly classifying an impaired subject) and specificity (probability of correctly classifying a cognitively normal subject) for each ML run of the testing samples. The same statistics were generated for each run of the cutoff scores of the total FAQ and MMSE without the use of ML methods, and for the stepwise logistic regression. Figures 2 and 3 respectively show the receiver operating characteristic (ROC) curves for the FAQ and MMSE total scores without ML methods, as well as the performance of the best results using various ML algorithms. Table 2 shows the classification results of each ML method and of published criteria for total MMSE and FAQ scores. A number of strategies were used to select an optimal decision tree for clinical use. We ordered pruned decision tree rule-sets by their frequency of occurrence across the different ML methods and runs. We examined the cross-validation procedure of CART, which selects the best single decision tree for a specified number of runs; we repeated this procedure 10 times. Each time, CART selected the same best decision tree. We also ran forward-stepping logistic regression on the dependent variable, *Dementia Class*, against the independent variables of the FAQ attributes (F-to-enter = 0.4, F-to-remove = 0.2) to identify the attributes which made statistically significant independent contributions to prediction of dementia status. For the demographic attributes (Table 1), only age and sex

Table 2. Sensitivity and Specificity of each Screening test by algorithm and published scoring criteria

FAQ (Normal = 137, Impaired = 211)

	CART	C45	C45Rules	FOCL	Naive Bayes	FAQ >8	FAQ >4
% Sensitivity	93	92	89	94	67	20	49
% Specificity	80	78	79	80	97	99	96
% Accuracy	88	88	85	89	83	51	68

MMSEPLUS (Normal = 198, Impaired = 227)

	C45	C45Rules	FOCL	Naive Bayes	MMSE >24	MMSE >27
% Sensitivity	77	70	79	66	30	62
% Specificity	80	86	70	87	100	81
% Accuracy	79	77	75	75	63	71

Ishihara Color Plates (Normal = 133, Impaired = 179)

	C45	C45Rules	NAIVE BAYES
% Sensitivity	68	68	73
% Specificity	55	52	52
% Overall	66	63	64

showed statistically significant differences between normal and impaired subjects. However, the age difference between normal and impaired subjects was less than one year, which is not a clinically significant difference. Therefore, only gender showed a clinically significant difference, with a preponderance of females in the normal group. The ICP attributes with ML methods resulted in at best a 73% sensitivity (52% specificity) using the Naive Bayes method and were not considered further. For the FAQ test, figure 2 shows that the FAQ with ML methods out-performed the best of the published cutoff criteria for the total FAQ score. It is interesting to note that the cutoff score of 9 or higher, recommended by the AHCPR, has a considerably poorer sensitivity for discriminating very mildly demented from normal subjects (20%) than that obtained for the ML methods, FOCL, C4.5, C4.5Rules, and CART (93%). One should also note that the number of questions needed to achieve these results with ML methods is markedly reduced. In the case of CART, only one question is required (*"Do you require assistance remembering appointments, holidays, family occasions, or taking medications?"*). For the MMSEPLUS test, figure 3 shows that, when used with ML methods, classification accuracy is always higher than when any total MMSE score is used as a cutoff criterion without ML methods. Using constrained vs. unconstrained analysis of the data with FOCL, there did not appear to be a significant improvement in classification accuracy, but no nonsense rules were generated when constraining FOCL with domain-specific knowledge. Given the various search strategies for finding the best decision tree or rule-set for clinical use, all approaches converged on one main conclusion: the response to a single question from the FAQ test gave classification accuracy as good as any other rule set and better than any published criteria. This question, *"Do you require assistance remembering appointments, family occasions, holidays or tak-*

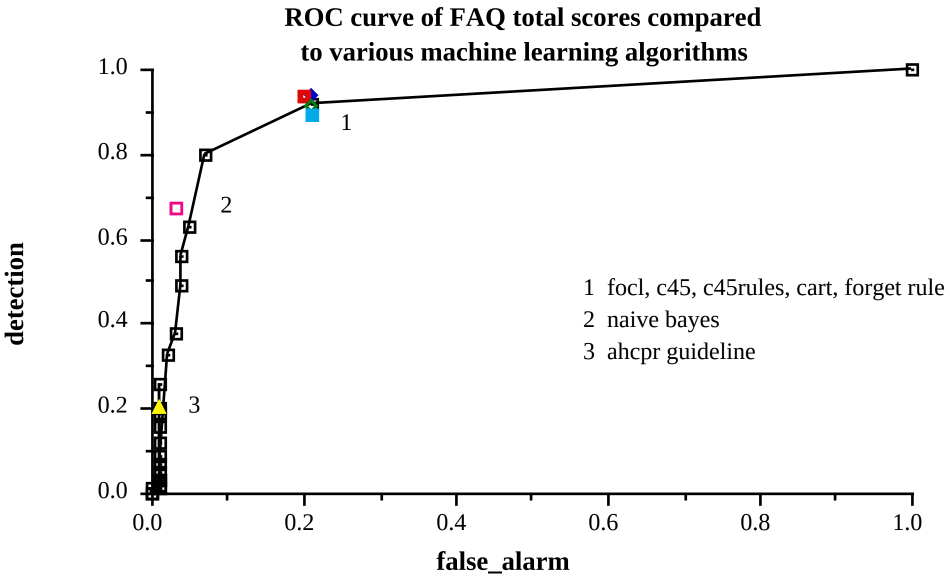

Fig. 2. FAQ ROC

ing medications?", we call the **forgetting rule**. All runs for all ML algorithms studied included this rule in the decision tree/rule-set; no other attribute was included in every decision tree/rule-set. Using CART's cross-validation procedure, this single rule decision tree was selected as the best tree on 10 out of 10 runs. Finally, forward-stepping logistic regression was used to identify the most important attributes in each run. These attributes were compared to those selected by the ML methods. For the FAQ, the forgetting attribute had the largest odds ratio on 32 of 33 runs (11.9 ± 7.6), and was the only attribute included in all 33 models. Job performance (odds ratio $= 4.2 \pm 4.2$) was the 2nd most frequently selected attribute, occurring in 20 of 33 runs. For the MMSEPLUS, the attribute, *# of trials to obtain 2 correct repetitions*, had overwhelmingly the highest odds ratio (90 ± 59), and was the first attribute entered for all 33 runs. The only other attribute included in all 33 runs was the *delayed recall attribute* ($1 \div oddsratio = 2.3 \pm 0.4$).

4 Discussion

There are four main findings of the present analysis. *First*, the ML methods can be interfaced with an electronic medical record system to learn directly from the data. The feasibility of this is also demonstrated by the work described in for example [2] and [14]. This feature contrasts with that of knowledge-based

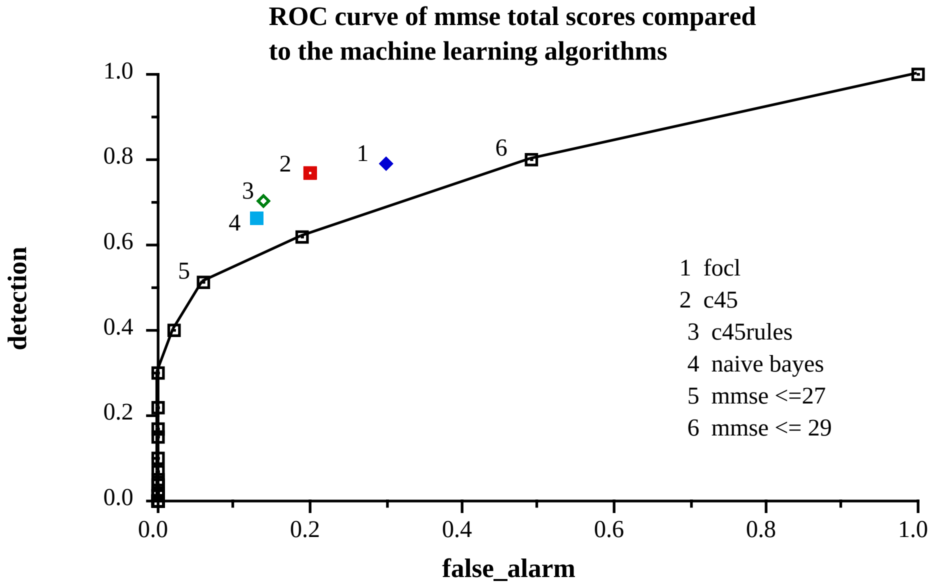

Fig. 3. MMSE ROC

systems, in which human experts design the decision rules and then test the data. Whereas humans usually select a few rules by which they make decisions, a machine can consider a larger number of rules. This is a specific advantage of ML methods. When supplemented by a review of the ML-generated rules or by incorporation of domain-specific knowledge into the ML algorithm, specific rules that violate domain knowledge can be minimized, thus enhancing the power of ML methods. This approach also identifies subtle logical errors in the electronic medical record that could be overlooked. For example, after reviewing a nonsense rule using job performance as a criterion, we discovered that some normal subjects had misinterpreted the question about their ability to perform a job, answering that they could no longer perform their job because they had retired. In fact, they were fully able to perform their job given the need to do so. The inconsistency in the attribute values was discovered, and corrected. Re-running the ML algorithm verified that the nonsense rule had been eliminated by this correction of the data. The *second* important finding of this paper is that ML methods used in conjunction with the MMSEPLUS test attributes outperform any published criteria for using total MMSE score to classify normal and cognitively impaired or very mildly demented subjects. They also do much better than any cutoff possible using the ROC curve. This supports the idea that some attributes of the MMSEPLUS are more important than others, and that the less important attributes may actually confuse classification. The findings of the logistic regression analyses did not substantially alter these conclusions. Two attributes, trials to learn address and recall of address performed as well as any other combination of MMSEPLUS attributes. The *third* important finding of this paper is clinical: when used with ML methods, a single question from the FAQ (the forgetting question) classifies normal cognitive and the mildest stages of a dementia as well as or better than any other combination of attributes from the FAQ, the MMSEPLUS and the ICP with and without total score, and outperforms any of the recommended scoring criteria for the FAQ or the MMSE total scores. The results of the logistic regression analyses confirmed the importance of the forgetting question. For screening purposes, we think that the tradeoff for higher sensitivity is preferable given the ease and applicability of the forgetting attribute as a screening test.

It is interesting to note that the AHCPR-recommended criteria for impairment using a total FAQ score of 9 or higher, is much higher than the score of a person answering positively only to the forgetting question (their FAQ total = 1-3 in that case). The higher total FAQ score recommended by the AHCPR is based on studies which included all levels of dementia severity. Using this criterion for the very mildly demented subjects in the present study resulted in only a 20% sensitivity, which implies that responses to other questions of the FAQ actually reduce the sensitivity for detecting very mild stages of dementia (compared to the forgetting rule alone). This is why inclusion of the total FAQ score as an attribute in the ML runs reduced the specificity and sensitivity when compared with the results obtained from analyses of the FAQ item attributes alone. The FAQ attributes therefore contribute unequally to dementia classification, with the forgetting question being the most contributory. This is our *fourth* significant finding.

4.1 Limits on Accuracy

Sample bias: The only demographic variable which differed to a clinically significant extent between normal and cognitively impaired subjects was gender. Since the decision rule sets rarely included gender in any of the ML runs and methods, we conclude that the findings presented here are not due to sample biases in age, education or gender. The findings are, however restricted to the population represented, which consists of individuals, mostly over 65 years and with more than a high school education. However, previous studies showing the insensitivity of the FAQ to educational level suggests that the results of this study apply to persons 65 or over, regardless of education.

5 Conclusions

We have successfully applied ML methods to increase sensitivity and specificity of commonly used dementia screening tests plus reduce the information required to make this decision. Additionally, ML methods can identify subtle errors in the electronic medical record which are due to misinterpretation of what is being asked of the subject. The rule set derived from the full data can be used on paper or as software in various clinical settings to enhance the detection of very early stages of a dementing illness. This should result in less disability per patient and better quality of life for both caregiver and patient through early intervention. The utility of ML-derived protocols with some human supervision has general applicability to many important medical areas, including cancer, heart disease, and stroke.

5.1 Acknowledgements

We thank professor Carl Cotman for helping establish a working relation with the AHCPR. This work was supported by the Alzheimer's Association Pilot Research Grant, PRG-95-161, *The Alzheimer's Intelligent Interface: Diagnosis, Education and Training.*

References

1. American Psychiatric Association, Washington, D. C. *Diagnostic and Statistical Manual of Mental Disorders*, 4 edition, 1994.
2. Ohmann C, Yang Q, Moustakis V, Lang K, and PJ van Elk. Machine learning techniques applied to the diagnosis of acute abdominal pain. In Pedro Barahona and Mario Stefanelli, editors, *Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine AIME95*, volume 934, pages 276–281. Springer, 1995.
3. Cestnik G, Konenenko I, and Bratko I. Assistant-86: A knowledge-elicitation tool for sophisticated users. In Bratko I and Lavrac N, editors, *Progress in Machine Learning*, pages 31–45. Sigma Press, 1987.
4. Crum R.M, Anthony J.C, Bassett S.S, and Folstein M.F. Population-based norms for the mini-mental state examination by age and educational level. *JAMA*, 269(18):2386–2390, May 1993.
5. Heckerman D.E, Horvitz E.J, and Nathwani B.N. Towards normative expert systems: Part i the pathfinder project. *Methods of Information in Medicine*, (31):90–105, 1992.
6. Duda R.O and Hart P.E. *Pattern Classification and Scene Analysis.* John Wiley, New York, 1973.
7. Ernst R.L and Hay J.W. The u.s. economic and social costs of alzheimer's disease revisited. *American Journal of Public Health*, 84(8):1261–4, Aug 1994.
8. Fillenbaum G.G, Heyman A, Wilkinson W.E, and Haynes C.S. Comparison of two screening tests in alzheimer's disease—the correlation and reliability of the mini-mental state examination and the modified blessed test. *Archives of Neurology*, 44(9):924–7, Sep 1987.
9. Folstein M.F, Folstein S.E, and McHugh P.R. Mini-mental state–a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–98, Nov 1975.
10. Hoffman R.S. Diagnostic errors in the evaluation of behavioral disorders. *JAMA*, 248:225–8, 1982.
11. Shinobu Ishihara. *Ishihara Tests for Colour-Blindness.* Kanehara Shuppan, Ltd., Tokyo Japan, 1994.
12. Wyatt J. Lessons learned from the field trials of acorn, a chest pain advisor. In Barber B, Cao D, Qin D, and Wagner F, editors, *Proceedings MedInfo*, pages 111–115. Elsevier Scientific, 1989.

13. Brieman L, Friedman J.H., Olshen R.A., and Stone C.J. *Classification and Regression Trees.* Wadsworth, Belmont, 1984.
14. Gierl L. and Stengel-Rutkowski S. Integrating consultation and semi-automatic knowledge acquisition in a prototype-based architecture: Experiences with dysmorphic syndromes. *Artificial Intelligence in Medicine*, 6:29–49, 1994.
15. Nada Lavrac and Igor Mozetic. Second generation knowledge acquisition methods and their application to medicine. In Keravnou E, editor, *Deep Models for Medical Knowledge Engineering*, pages 177–198. Elsevier, New York, 1992.
16. Lubeck D.P, Mazonson T and Bowe P.D. Potential effect of tacrine on expenditures for alzheimer's disease. *Medical Interface*, 7(10):130–8, Oct 1994.
17. McCleary R, Shankle W.R, Mulnard R.A, and Dick M.B. Ishihara test performance and dementia. *Journal of the Neurological Sciences*, in press 1996.
18. R.S. Michalski, Mozetic I, Hong J, and Lavrac N. The multi-purpose incremental learning system aq15 and its testing application to three medical domains. In *In Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 1041–1045, Philadelphia, PA, 1986. Morgan Kaufmann.
19. Morris J.C. The clinical dementia rating (cdr): current version and scoring rules. *Neurology*, 43(11):2412–4, Nov 1993.
20. Igor Mozetic and Bernhard Pfahringer. Improving diagnostic efficiency in kardio: Abstractions, constraint propagation and model compilation. In Keravnou E, editor, *Deep Models for Medical Knowledge Engineering*, pages 1–25. Elsevier, New York, 1992.
21. O'Connor D.W, Fertig A, Grande M.J, Hyde J.B, Perry J.R, Roland M.O, Silverman J.D and Wright S.K. Dementia in general practice: the practical consequences of a more positive approach to diagnosis. *Br J Gen Pract*, 43:185–8, 1993.
22. Oconnor D.W, Pollitt PA, Treasure F.P, Brook C.P.B, and Reiss B.B. The influence of education, social class and sex on mini-mental state scores. *Psychological Medicine*, 19:771–776, 1989.
23. Michael Pazzani and Dennis Kibler. The utility of knowledge in inductive learning. *Machine Learning*, (9):57–94, 1992.
24. Pfeffer R.I, Kurosaki T.T, Harrah C.H, Chance J.M, and Filos S. Measurement of functional activities in older adults in the community. *J Gerontology*, 37:323–9, 1982.
25. Quinlan J.R. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, Los Altos, California, 1993.
26. Kohavi R, George John, Richard Long, David Manley, and Karl Pfleger. Mlc++: A machine learning library in c++. In *Tools with Artificial Intelligence*, pages 740–743. IEEE Computer Society Press, 1994.
27. Shankle W.R, Datta P, Dillencourt M, and Pazzani M. Improving dementia screening tests with machine learning methods. *Alzheimer's Research*, 2(3), Jun 1996.
28. Shortliffe E. *Computer-Based Medical Consultations: MYCIN.* Elsevier/North Holland, New York, 1976.
29. Welsh K.A, Butters N, Mohs R.C, Beekly D, Edland S, and Fillenbaum G. The consortium to establish a registry for alzheimer's disease (cerad. part v. a normative study of the neuropsychological battery. *Neurology*, 44(4):609–14, Apr 1994.
30. Williams T.F and Costa P.T. Recognition and initial assessment of alzheimer's disease and related dementias: Clinical practice guidelines. Technical report, Department of Health and Human Services, 1995.

This article was processed using the LaTeX macro package with LLNCS style