

## ERP Based Decision Fusion for AD Diagnosis Across Cohorts

Metin Ahiskali, Deborah Green, John Kounios, Christopher M. Clark and Robi Polikar\*

**Abstract**—As the average life expectancy increases, particularly in developing countries, prevalence of neurodegenerative diseases has also increased. This trend is especially alarming for Alzheimer’s disease (AD); as there is no cure to stop or reverse the effects of AD. However, recent pharmacological advances can slow the progression of AD, but only if AD is diagnosed at early stages. We have previously introduced an ensemble of classifiers based approach for combining event related potentials obtained from different electrode locations as an effective approach for early diagnosis of AD. We further expand this approach and analyze its robustness and stability in two ways: comparing the diagnostic accuracy on hand selected and cleaned data vs. standard automated preprocessing, but more importantly, comparing the diagnostic accuracy on two different cohorts, whose data are collected under different settings: a research university lab and a community clinic.

### I. INTRODUCTION

Alzheimer’s disease is the fifth leading cause of death among the elderly, and the sixth leading cause of death among all age groups in the US. Given that there is no treatment available to stop or reverse the effects of AD, and no definitive method for diagnosing the disease with certainty, combined with the availability of recently developed medications for slowing the progression of the disease, it has become increasingly important to diagnose AD at its earliest stages possible.

A neurodegenerative disease, AD is characterized by neuronal death, leading to declines in memory, motor skills and cognitive ability. While two misfolded proteins,  $\beta$ -amyloid and hyperphosphorylated- $\tau$ , which cause plaques and neurofibrillary tangles, respectively, have been implicated as causing AD, the genesis of these proteins is unknown. Perhaps more perplexing, recent studies have shown that removal of  $\beta$ -amyloid from the brain using targeted vaccination does not reverse the damage caused [1]. With overall life expectancy increasing, more notably in developed countries, the number of AD cases has grown tremendously since its first discovery over a century ago. The vast majority of AD patients are over the age of 65, with an average 2% of cases affected under 65 (early onset AD), 19% between 75 and 84, and 42% ~50% above the age of 85. With over 5 million people suffering from AD alone, the enormous financial and emotional cost of the disease on the patient, family, and society in general has made AD a major health concern [2].

Currently, a definitive diagnosis of AD can only be made post-mortem by analyzing the brain tissue under the microscope for the presence of plaques and tangles, a method only available during an autopsy. More recently, cere-

brospinal fluid (CSF) concentrations of hyperphosphorylated- $\tau$  and  $\beta$ -amyloid proteins have been shown to be diagnostically informative biomarkers. However, such an analysis requires invasive, expensive, extremely painful – and hence undesirable – lumbar puncture, which also requires specialty clinics, research or university hospitals.

The most common method of pre-mortem diagnosis is a longitudinal clinical evaluation, performed over a year or longer, which includes several batteries of memory tests as well as interviews of the patient and their caretakers. The primary path of diagnosis using this approach is through documenting cognitive decline over time and eliminating all other possible disorders. Such an evaluation is therefore subjective, tedious, requires expert neuropsychologists, and due to large amount of time involved in the process, is extremely expensive. While such an evaluation can achieve over 90% accuracy in AD diagnosis, when conducted by expert neurologists at large university hospitals or dementia clinics, geographical and financial restrictions generally limit most patients to smaller hospitals and clinics, where the diagnostic accuracies are estimated to be around 75%, even with the benefit of longitudinal evaluations [3].

Given the prevalence of the disease, the expense and expertise needed for the less-than-perfect diagnosis, coupled with the availability of new drugs that work best if administered early, there is a compelling case for the urgent need for a reliable, inexpensive, non-invasive diagnostic tool that can be made available to community clinics, where most people seek medical help. The event related potentials (ERPs) obtained from the electroencephalogram (EEG) may provide just such a tool. Long overlooked due to its poor spatial resolution and difficulty of interpretation, recent advances in signal processing and machine learning have shown that early dismissal of ERPs as a potential biomarker were unfounded. In fact, several recent studies using the well-known oddball paradigm have demonstrated that the decreased amplitudes and increased latencies of the P300 component of ERPs – a positive peak that occurs around 300 ms after a stimulus is observed by the subject – is linked to cognitive decline [4-7]. Various signal processing approaches on the raw EEG or the P<sub>300</sub> has been conducted since then, verifying the presence of a statistical correlation, that showed limited success in patient specific diagnosis [8-10]. In our early work, we have shown that discrete wavelet coefficients of the ERPs, and not that of just the P<sub>300</sub> components, are quite beneficial in patient specific AD diagnosis, particularly when the ERPs in response to different types of stimuli are combined [11-14].

Previously, we introduced an ensemble of classifiers based decision fusion approach, showing the feasibility of combining ERPs acquired from different electrodes for automated early diagnosis [13;15]. In this contribution, we extend our analysis to include the robustness of the approach with respect to manual hand cleaning vs. automated preprocessing, but more importantly with respect to different cohorts. To the best of our knowledge, this is the first study of its kind that compares the performance on different cohorts whose data are acquired at different centers.

---

This work is supported by NIH grants AG022272 and AG10124, PA Dept. of Health, SAP4100027296, and by NSF grant no ECS 0239090.

M. Ahiskali and R. Polikar are with Signal Processing and Pattern Recognition Laboratory, Dept. of Electrical and Computer Eng. at Rowan University, Glassboro, NJ 08028, USA. (e-mail: mahiskali@gmail.com).

C.M. Clark is with the Department of Neurology, University of Pennsylvania, Philadelphia, PA (e-mail: chris.clark@uphs.upenn.edu).

D. Green and J. Kounios are with the Dept. of Psychology, Drexel University, Philadelphia, PA (e-mail: dgreen@drexel.edu; Kounios@drexel.edu)

\*Contact author: Robi Polikar, phone: (856) 256 – 5372; fax: (856) 256-5241. E-mail: polikar@rowan.edu.

## II. EXPERIMENTAL SETUP

### A. Patient Cohorts

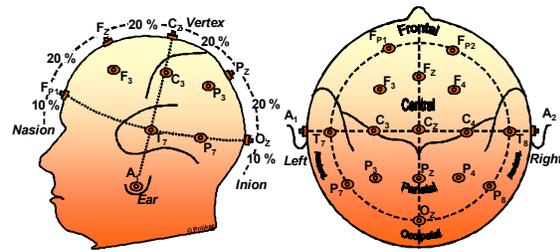
Two different cohorts were recruited, which we will refer to as Cohort A and Cohort B in the rest of this paper. Cohort A subjects were primarily recruited between 2003 and 2006, with particular emphasis on AD subjects being in the earliest stages of the disease. Cohort B subjects were recruited between 2006 and 2009, where AD group represented a wider spectrum of disease stages. While there are many indicators to determine the severity of the disease, one of the simpler and effective tests is the Mini Mental State Exam (MMSE). The MMSE is a standardized test performed to access the subject's attention, immediate and short-term memory, language, as well as their ability to follow various written and verbal communication. For healthy individuals over the age of 80, the median score is a 25 out of possible 30 total points (for no cognitive decline); however, this score is strongly proportional to number of years of education. The median MMSE is 29, for those with at least 9 years of schooling, and 22 for those with zero to four years of education. A score of 19 or lower is a clear indicator of cognitive impairments and hence AD pathology [16]. Cohort A, had 34 AD patients with  $\mu_{AGE(AD)}=74$ ,  $\mu_{MMSE(AD)}=25$ , and 37 controls (CN) with  $\mu_{AGE(CN)}=76$ , and  $\mu_{MMSE(CN)}=29$ , indicating that the AD patients were in their earliest stages of the disease. For Cohort B, we had 27 AD patients with  $\mu_{AGE(AD)}=77$ ,  $\mu_{MMSE(AD)}=23$ , and 26 control subjects with  $\mu_{AGE(CN)}=76$ , and  $\mu_{MMSE(CN)}=29$ .

Cohort B subjects continue to be recruited, and over 100 subjects have been recruited so far. However, unlike Cohort A, Cohort B data acquisition is done – not in a carefully controlled research laboratory setting, but rather in a setting that mimics a community clinic. Hence the data are substantially noisier than those for Cohort A. Therefore, for this specific test, we chose a small subset of these patients, whose data were manually and carefully analyzed and cleaned by a specially trained EEG technician. We compared the performance of the proposed system on the identical subjects whose data were manually cleaned by the technician vs. automatically cleaned by a preprocessing scheme. Furthermore, while the Cohort A EEG data were obtained using 19 electrodes, Cohort B data were acquired using only 16 electrodes (using a simpler amplifier, mimicking one that may be available to community clinics).

For both Cohort A and B, inclusion criteria for AD group was satisfying the NINCDS-ADRDA criteria [17] for probable AD, which includes a battery of memory tests (including MMSE), interviews with the subject and their caregivers, clinical dementia rating score of 0.5 or higher for AD cohort and 0 for the normal cohort. All subjects were over 60 years old. Exclusion criteria for both groups were evidence of any other central nervous system damage or use of sedatives, anxiolytic or antidepressants within 48 hours of ERP acquisition.

### B. The Oddball Paradigm and ERP Acquisition

The ERPs acquired from the patients were collected using an auditory oddball paradigm protocol. Electrodes were placed according to 10-20 standard (Fig 1), whose impedances were kept below 20 $\Omega$ . Each subject was tested for 30 minutes with approximately three minutes of rest for every five minutes of testing. A 1 kHz tone was presented to both ears, 60 dB above each subjects hearing threshold.



Decision fusion based data fusion refers to combination of decisions made by individual classifiers, as opposed to feature based data fusion where individual features are usually concatenated. We use a modified (augmented) stacked generalization approach to obtain feature level decision fusion, followed by a combination of such decisions to create the final across-features decision fusion.

### C. Stacked Generalization

The primary goal of stacked generalization is to confirm or correct what has been learned by a group of preliminary (Tier-1) classifiers with the use of a meta-classifier. Any instance in a certain region of the feature space (e.g., near the decision boundary), may be more likely to be misclassified by certain classifiers than others. This trend can be learned by mapping the outputs of an ensemble of classifiers to their true labels (Fig 2.), thereby allowing the system to adapt to classifiers with varying performances.

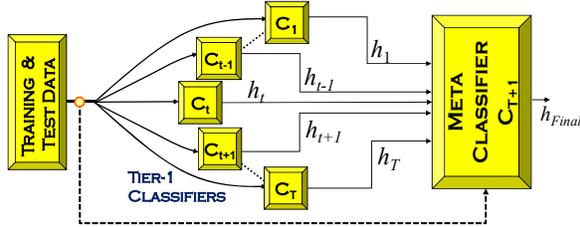


Fig. 2 - Stacked Generalization block diagram

### D. Augmented Stacked Generalization

In modifying the basic stacked generalization algorithm, we augment the outputs of the Tier-1 classifiers by concatenating the *original data* before training the meta-classifier (dashed line in Fig. 2), which we refer to as augmented stacked generalization (ASG). This process enriches the intermediate feature space used by the meta-classifier, which in turn aids in overall system performance. We implement a two stage training process, a leave-one-out (LOO) nested into a 5-fold cross validation. For each fold held out as test data, the remaining 4 folds are used in a LOO fashion to train the classifiers:  $N-1$  instances are used to train the Tier-1 classifiers and the outputs of the Tier 1 classifiers on the remaining  $N^{\text{th}}$  instance is used for training the meta-classifier. The system is then tested on the test block of the 5-fold cross validation. For each feature set (ERPs from different electrode locations, frequency bands and stimulus types), the entire process is then repeated 10 times, creating a different partition of 5 folds. Upon completion, we obtain a different ASG based system trained on different specific feature set (e.g. NP<sub>Z</sub>12—Novel stimuli, P<sub>Z</sub> electrode, 1-2 Hz frequency band). This constitutes *one* decision fusion based expert for its respective feature set. Such experts are then combined using sum, and simple majority voting to obtain final decision fusion classification, combining different feature sets. Figure 3 illustrates the overall system diagram.

Multilayer perceptron based classifiers were used throughout the entire system, since these classifiers naturally provide an output that can be interpreted as a support for each class. Nine MLPs with 10 hidden layer nodes were used in Tier 1, and one MLP with 30 hidden layer nodes was used as the meta-classifier. All MLPs were trained with conjugate gradient based backpropagation, with a momentum term of 0.95 and an error goal of 0.05.

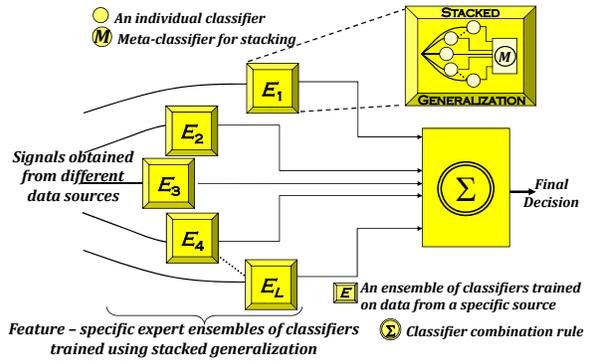


Fig. 3 – Overall system block diagram

## IV. RESULTS

Since Cohort A and B utilize different electrodes, not all feature sets can be compared directly between the two. However, viable substitutions were made (see Tables 1 and 3 for feature set lists for each cohort), and the identical training and testing procedure was used for both.

Recall that Cohort B data underwent two different methods for cleaning and artifact rejection. First, it was *hand cleaned* by a trained EEG technician, manually removing or compensating for each artifact (e.g., muscular movement, eye blinks, etc.). In the second method, we employed a simple automated artifact rejection using derivative based thresholding and baselining. The analysis was then repeated on both sets of data to determine the robustness of the decision fusion based analysis with respect to different artifact rejection methods. This is because; the proposed automated system is intended for community clinics. Therefore, two performance metrics are given for all Cohort B calculations, each corresponding to either hand cleaning of the data or automated artifact rejection.

### A. Cohort A Diagnostic Accuracies

The diagnostic accuracies for the highest performing 16 individual feature sets for Cohort A are shown in Table 1. The averages and 95% confidence intervals are based on 10 independent trials of 5-fold cross validation.

TABLE 1 – COHORT A: INDIVIDUAL FEATURE SET PERFORMANCES

FS	Avg (%)	Best Trial	FS	Avg (%)	Best Trial
NP <sub>Z</sub> 12	70.1±1.2	88.6	NT <sub>8</sub> 12	64.0±1.6	79.1
NP <sub>Z</sub> 01	69.4±1.8	77.1	TF <sub>p2</sub> 12	63.0±1.2	73.7
NC <sub>Z</sub> 12	69.1±1.4	78.0	NP <sub>Z</sub> 24	62.6±1.8	79.7
TP <sub>Z</sub> 12	67.5±1.6	81.0	TP <sub>Z</sub> 24	62.5±1.0	81.3
TP <sub>Z</sub> 24	66.9±0.9	89.9	TP <sub>Z</sub> 01	62.5±1.9	79.7
TF <sub>8</sub> 12	66.5±1.9	82.5	TP <sub>3</sub> 12	62.4±1.2	79.5
NF <sub>Z</sub> 24	66.2±1.7	75.1	TC <sub>Z</sub> 24	61.8±1.7	79.4
NC <sub>Z</sub> 24	64.7±1.2	88.2	NO <sub>Z</sub> 12	64.0±1.6	79.1

The better performing individual feature sets (FS) were then combined to complete the decision fusion based ensemble system. We chose three such combinations:

**FS1:** NC<sub>Z</sub>12+NC<sub>Z</sub>24+NP<sub>Z</sub>24+NT<sub>8</sub>12+NP<sub>Z</sub>24+TFP<sub>Z</sub>12+TP<sub>Z</sub>01+TF<sub>8</sub>12+TP<sub>Z</sub>24

**FS2:** NP<sub>Z</sub>12+NP<sub>Z</sub>24+NC<sub>Z</sub>24+NT<sub>8</sub>12+TF<sub>p2</sub>12+NC<sub>Z</sub>12

**FS3:** TF<sub>p2</sub>12+TF<sub>8</sub>12+TP<sub>Z</sub>01+NC<sub>Z</sub>12+NP<sub>Z</sub>24+NT<sub>8</sub>12+NP<sub>Z</sub>12

The results of the combinations are shown in Table 2 for sum and simple majority voting (SMV) based combination rules, which shows that the sum rule in general performing better, with statistical significance.

TABLE 2 – COHORT A: DECISION FUSION DIAGNOSTIC ACCURACIES

SUM		Avg (%)	Best	SMV		Avg(%)	Best
	FS1	84.5±1.8	94.7		FS1	74.3±1.3	82.1
	FS2	81.3±1.7	83.0		FS2	71.8±2.1	86.6
	FS3	82.4±2.0	87.0		FS3	73.1±2.2	82.0

### B. Cohort B Diagnostic Accuracies

Diagnostic accuracies for each feature set are shown in Table 3, for automated and hand cleaned preprocessing.

TABLE 3– COHORT B: INDIVIDUAL FEATURE SET PERFORMANCES

Feature Set	Computer		Hand-Cleaned	
	Avg (%)	Best Trial	Avg (%)	Best Trial
NPZ12	62.9±1.5	90.0	63.8±2.55	90.0
NT812	65.8±2.4	80.0	59.3±1.72	80.0
NPZ24	59.3±0.9	70.0	60.9±0.80	76.9
NCZ24	62.7±1.2	80.0	55.6±1.49	80.0
NCZ12	65.6±2.5	90.0	65.7±3.22	80.0
NOZ12	58.8±2.3	90.0	58.9±1.61	80.0
TCZ24	60.5±2.0	90.0	60.6±2.01	90.0
TPZ12	57.8±1.1	70.0	61.0±2.05	90.0
TP312	57.5±2.3	80.0	61.4±1.60	80.0
NFZ24	60.9±2.4	90.0	59.5±1.94	80.0
TP324	65.8±2.2	90.0	62.0±1.11	90.0
TPZ24	65.8±1.3	84.6	58.5±2.36	80.0
NPZ01	59.5±1.4	84.6	64.2±1.37	80.0
TF312	58.7±1.6	90.0	68.4±1.15	80.0
TFP212	65.6±1.2	90.0	63.6±2.34	90.0
TPZ01	62.9±2.7	100.0	57.5±0.82	70.0

Since different electrodes were used in Cohort B, the FS combinations vary slightly from those of Cohort A, with the only substitution being TP<sub>3</sub>12 replacing TF<sub>8</sub>12. Results of the final combinations, shown with each combination rule are shown in Table 4 for automated preprocessing, and in Table 5 for hand-cleaning based preprocessing.

TABLE 4– COHORT B: DECISION FUSION DIAGNOSTIC ACCURACIES BASED ON AUTOMATED PREPROCESSING

SUM		Avg (%)	Best	SMV		Avg (%)	Best
	FS1	76.04±1.2	81.12		FS1	72.26±1.2	82.83
	FS2	79.11±1.8	86.78		FS2	71.88±1.9	79.15
	FS3	83.13±1.6	96.23		FS3	76.91±0.9	86.14

TABLE 5– COHORT B: DECISION FUSION DIAGNOSTIC ACCURACIES BASED ON HAND-CLEANED PREPROCESSING

SUM		Avg (%)	Best	SMV		Avg (%)	Best
	FS1	73.05±0.9	80.00		FS1	72.28±0.7	79.91
	FS2	78.16±1.9	83.45		FS2	73.52±1.2	77.21
	FS3	83.00±2.1	94.99		FS3	74.11±1.9	76.55

## V. CONCLUSIONS

We make several observations from the results presented above. First, ASG based ensemble performance for any of the feature sets exceeds that of both community clinic diagnostic accuracy, as well as individual feature set performances (in Tables 1 and 3), demonstrating the feasibility of the approach. Second, the performance differences between Cohort A and Cohort B are not significant, nor are the differences between hand-cleaned preprocessing and automated preprocessing. This is important, indicating the ASG based decision fusion approach - when used with wavelet coefficients as features - is quite robust with respect to cohort differences as well preprocessing schemes. Third, the performance differences among the top three feature sets are all also very similar, with FS3 performing slightly better (with significance only in Cohort B), indicating that several different electrode combinations may be used for the analysis. The sum rule, however, performed significantly better than SMV across most comparisons.

Overall, we conclude that i) the approach is feasible as a first screening diagnosis tool to be used at community clinics, ii) there is in fact complementary information in ERPs obtained from different electrodes, which can be fused through decision fusion, and iii) the approach is quite robust to cohort and preprocessing variations. Future work will include expanding the analysis to a larger cohort, and training on one cohort and evaluating on the other.

## REFERENCES

- [1] S. Gilman, M. Koller, R. S. Black, and L. Jenkins, "Clinical effects of Aβ immunization (AN1792) in patients with AD in an interrupted trial," *Neurology*, vol. 64, pp. 1563-1572, 2005.
- [2] Alzheimer's Assoc., 2009AD Facts and Figures, Available online: [http://www.alz.org/national/documents/report\\_alzfactsfigures2009.pdf](http://www.alz.org/national/documents/report_alzfactsfigures2009.pdf)
- [3] A. Lim, D. et al., "Clinico-neuropathological correlation of Alzheimer's disease in a community-based case series," *J Am Geriatr Soc*, vol. 47, no. 5, pp. 564-569, May1999.
- [4] S. Yamaguchi, H. Tsuchiya, S. Yamagata, G. Toyoda, and S. Kobayashi, "Event-related brain potentials in response to novel sounds in dementia," *Clinical Neurophys.*, vol. 111, no. 2, pp. 195-203, 2000.
- [5] D. J. Linden, "The p300: where in the brain is it produced and what does it tell us?," *Neuroscientist*, vol. 11, no. 6, pp. 563-576, 2005.
- [6] K. Bennys, F. Portet, J. Touchon, and G. Rondouin, "Diagnostic Value of Event-Related Evoked Potentials N200 and P300 Subcomponents in Early Diagnosis of Alzheimer's Disease and Mild Cognitive Impairment," *J. Clinical Neurophys.*, vol. 24, no. 5, pp. 405-412, 2007.
- [7] J. Polich, "Updating P300: An integrative theory of P3a and P3b," *Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128-2148, Oct.2007.
- [8] M. Lindau, V. Jelic, S. E. Johansson, C. Andersen, L. O. Wahlund, and O. Almkvist, "Quantitative EEG abnormalities and cognitive dysfunctions in frontotemporal dementia and Alzheimer's disease," *Dement. Geriatr. Cogn Disord.*, vol. 15, no. 2, pp. 106-114, 2003.
- [9] M. Karrasch, M. Laine, J. O. Rinne, P. Rapinaja, E. Sinerva, and C. M. Krause, "Brain oscillatory responses to an auditory-verbal working memory task in mild cognitive impairment and Alzheimer's disease," *Int. Journal of Psychophysiology*, vol. 59, no. 2, pp. 168-178, 2006.
- [10] T. Demiralp, A. Ademoglu, Y. Istefanopoulos, C. Basar-Eroglu, and E. Basar, "Wavelet analysis of oddball P300," *International Journal of Psychophysiology*, vol. 39, no. 2-3, pp. 221-227, Jan.2001.
- [11] G. Jacques, J. L. Frymiare, J. Kounios, C. Clark, and R. Polikar, "Multiresolution analysis for early diagnosis of Alzheimer's disease," *IEEE Engineering in Medicine and Biology Conference (EMBC 2004)*, vol. 1, 2004, pp. 251-254.
- [12] D. Parikh, N. Stepenosky, A. Topalis, D. Green, J. Kounios, C. Clark, and R. Polikar, "Ensemble Based Data Fusion for Early Diagnosis of Alzheimer's Disease," 2005, pp. 2479-2482.
- [13] R. Polikar, A. Topalis, D. Parikh, D. Green, J. Frymiare, J. Kounios, and C. M. Clark, "An ensemble based data fusion approach for early diagnosis of Alzheimer's disease," *Information Fusion*, vol. 9, no. 1, pp. 83-95, Jan.2008.
- [14] R. Polikar, A. Topalis, D. Green, J. Kounios, and C. M. Clark, "Comparative multiresolution wavelet analysis of ERP spectral bands using an ensemble of classifiers approach for early diagnosis of Alzheimer's disease," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 542-558, Apr.2007.
- [15] T. Patel, R. Polikar, C. Davatzikos, and C. M. Clark, "EEG and MRI data fusion for early diagnosis of Alzheimer's disease," *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, 2008, pp. 1757-1760.
- [16] R. M. Crum, Anthony J.C., Bassett S.S., and M. Folstein, "Population-based norms for the Mini-Mental State Examination by age and educational level," *Journal of American Medical Association*, vol. 269, no. 18, pp. 2386-2391, 1993.
- [17] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease," *Neurology*, vol. 34, no. 7, pp. 939-944, 1984.
- [18] R. Polikar, "Bootstrap - Inspired Techniques in Computation Intelligence," *Signal Processing Magazine, IEEE*, vol. 24, no. 4, pp. 59-72, 2007.