Available online at www.sciencedirect.com

ScienceDirect

Information Fusion 9 (2008) 83–95

INFORMATION FUSION

www.elsevier.com/locate/inffus

An ensemble based data fusion approach for early diagnosis of Alzheimer's disease

Robi Polikar [a,*], Apostolos Topalis [a], Devi Parikh [a], Deborah Green [b], Jennifer Frymiare [b], John Kounios [b], Christopher M. Clark [c]

[a] *Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ 08028, United States*
[b] *Department of Psychology, Drexel University, Philadelphia, PA 19102, United States*
[c] *Department of Neurology, University of Pennsylvania, Philadelphia, PA 19104, United States*

Received 15 December 2005; received in revised form 22 September 2006; accepted 29 September 2006
Available online 30 November 2006

Abstract

As the number of the elderly population affected by Alzheimer's disease (AD) rises rapidly, the need to find an accurate, inexpensive and non-intrusive diagnostic procedure that can be made available to community healthcare providers is becoming an increasingly urgent public health concern. Several recent studies have looked at analyzing electroencephalogram (EEG) signals through the use of wavelets and neural networks. While showing great promise, the final outcomes of these studies have been largely inconclusive. This is mostly due to inherent difficulty of the problem, but also – perhaps – due to inefficient use of the available information, as many of these studies have used a single EEG channel for the analysis. In this contribution, we describe an ensemble of classifiers based data fusion approach to combine information from two or more sources, believed to contain complementary information, for early diagnosis of Alzheimer's disease. Our emphasis is on sequentially generating an ensemble of classifiers that explicitly seek the most discriminating information from each data source. Specifically, we use the event related potentials recorded from the Pz, Cz, and Fz electrodes of the EEG, decomposed into different frequency bands using multiresolution wavelet analysis. The proposed data fusion approach includes generating multiple classifiers trained with strategically selected subsets of the training data from each source, which are then combined through a modified weighted majority voting procedure. The implementation details and the promising outcomes of this implementation are presented.

Keywords: Data fusion; Multiple classifier/ensemble systems; Learn^{++}; Alzheimer's disease; Early diagnosis

1. Introduction

This paper is concerned with investigating the feasibility of an ensemble of classifiers based data fusion approach on a specific medical application; however, our goal is not merely presenting just an interesting application on which ensemble systems happen to work well. Rather, we try to make the case that the framework provided by the proposed ensemble based approach fits naturally to applications where data from different sources need to be combined. In other words, the attributes of the proposed approach closely match the characteristics of the underlying problem. The importance of such a close match is only amplified by the fact that the specific application we investigate is itself a significant public health concern, and that it has widespread impact on long term geriatric care.

1.1. Senile dementia of Alzheimer's type

Neurological disorders that cause gradual loss of cognitive function are collectively known as dementia. Among several forms of dementia, the most common form is the

[*] Corresponding author. Tel.: +1 856 256 5372; fax: +1 856 256 5241.
E-mail address: polikar@rowan.edu (R. Polikar).

1566-2535/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2006.09.003

irreversible and incurable senile dementia of Alzheimer's type, or just Alzheimer's disease (AD), in short. Once considered a rare disease, and mostly ignored due to elderly people being its primary victim, the number of people suffering from AD has been growing rapidly as the world's population ages. Today, it is estimated that there are 18 million people suffering from AD worldwide, two-thirds of whom live in developed or developing countries. This figure is expected to soar to 34 million by 2025. In the US alone, over 4.5 million (1.5% of the total population) suffer from AD, which is expected to reach 12–16 million by 2050. Up to age 60, AD appears in only 1% or less of the population, but its prevalence increases sharply, doubling every five years: the disease affects 5% of 65-year olds and 30–50% of 85-year olds [1,2].

Apart from its slow but debilitating effects on its victim, the disease has a devastating financial toll on the society (estimated at over $100 billion annually in the US alone), and causes an immeasurable grief on the victim's caregivers. While the specific causes of AD are unknown, the disease is characterized by abnormal proteins in the brain that comprise neurofibrillary tangles and plaques. These proteins can only be identified by examining the brain tissue under a microscope. Hence the only form of definitive diagnosis is an autopsy.

Several biomarkers have been linked to AD, such as the cerebrospinal fluid tau, β-amyloid, urine F2-isoprostane, and brain atrophy detected by PET/MRI scan. However, none of these methods has proven to be conclusive for early diagnosis, and even if they were, they remain primarily university and research based tools. Currently, clinical and neuropsychological evaluations achieve an average positive predictive value of 90%; however, this level of expertise is typically available only at university or research clinics, can be very expensive, and hence remain beyond reach for most patients. Therefore, these patients are evaluated by local community healthcare providers, where the expertise and accuracy of AD specific diagnosis remains uncertain. In fact, a recent study reported that, despite the advantage of longitudinal follow up, a group of Health Maintenance Organization based physicians had an overall accuracy of 75% for the clinical diagnosis of AD [3]. Meanwhile, active development of pathologically targeted medications requires an accurate diagnosis at the earliest stage possible. Only then can the patient's life expectancy and quality of life be improved significantly. To have a meaningful impact on healthcare, the diagnostic tool must be inexpensive, non-invasive, accurate, and available to community physicians.

Event related potentials (ERPs) of the electroencephalogram (EEG) may provide such a tool. However, the ability of EEG signals to resolve AD specific information is typically masked by changes due to normal aging, coexisting medical illness, and levels of anxiety or drowsiness during measurements. The ERPs of the EEG, obtained through the oddball paradigm protocol, has previously been linked to cognitive functioning, and is believed to be relatively insensitive to above-mentioned parameters. In this protocol, subjects are instructed to respond to an occasionally occurring target (oddball) stimulus, within a series of regular non-target stimuli. The ERPs then show a series of peaks, among which the P300 – a positive peak with an approximate latency of 300 ms that occurs only in response to the oddball stimulus – is of particular interest. Changes in the amplitude and latency of the P300 (P3, for short) are known to be altered by neurological disorders, including AD, that affects the temporal–parietal regions of the brain [4]. Specifically, increased latency and decreased amplitude of P300 is associated with AD [5–7]. However, looking at just the P300 component – while provides statistical correlation with AD – does not help in identifying individual patients: cognitively normal people may have delayed or absent P300; and those with AD, in particular early stages, may still have a strong P300, as shown in Fig. 1.

Traditional ERP analysis is performed either in time or frequency domain. However, both are individually suboptimal, since the ERP is a time and frequency varying signal. Despite its now mature history, studies applying time–frequency techniques, such as wavelets, to ERPs have only recently started, and mostly on non-AD related studies designed specifically for P300 analysis [8–12]. Studies directly targeting AD diagnosis using discriminant analysis on ERP features [13], or wavelet analysis followed by neural network classification have been recently tried by others [14–16], and ourselves [17,18]. These efforts have only achieved limited success, however, in part due to difficulty of the problem, in part due to lack of a large study cohort, and in part due to using single channel data from the EEG. Finally, combining various spectral components of the EEG from different channels [19,20], or combining EEG with other imaging modalities (such as MRI, MEG, etc.) [21] have also been tried for brain function analysis, but not for AD diagnosis.

In most P300 studies specifically designed for detecting AD induced changes, the ERPs are typically obtained from the Pz electrode [22] (of the 10–20 EEG electrode placement system, see Fig. 2), and only for responses to the oddball tones, where and when the P300 is known to be most prominent. However, we believe that the nearby electrodes, such as Cz and Fz, may also carry complementary information, even when the subjects hear – but do not respond – to novel tones. The question is then whether the additional signals do in fact carry complementary information, and if so, how such pieces of information can be best combined for improved diagnostic performance.

In this study, we describe an ensemble of classifiers based data fusion approach for this problem, where a separate ensemble of classifiers are trained with data from each source, and their outputs are combined through a modified weighted majority voting procedure. The procedure used for generating the ensemble of classifiers is the Learn^{++} algorithm. Learn^{++} was inspired in part by the

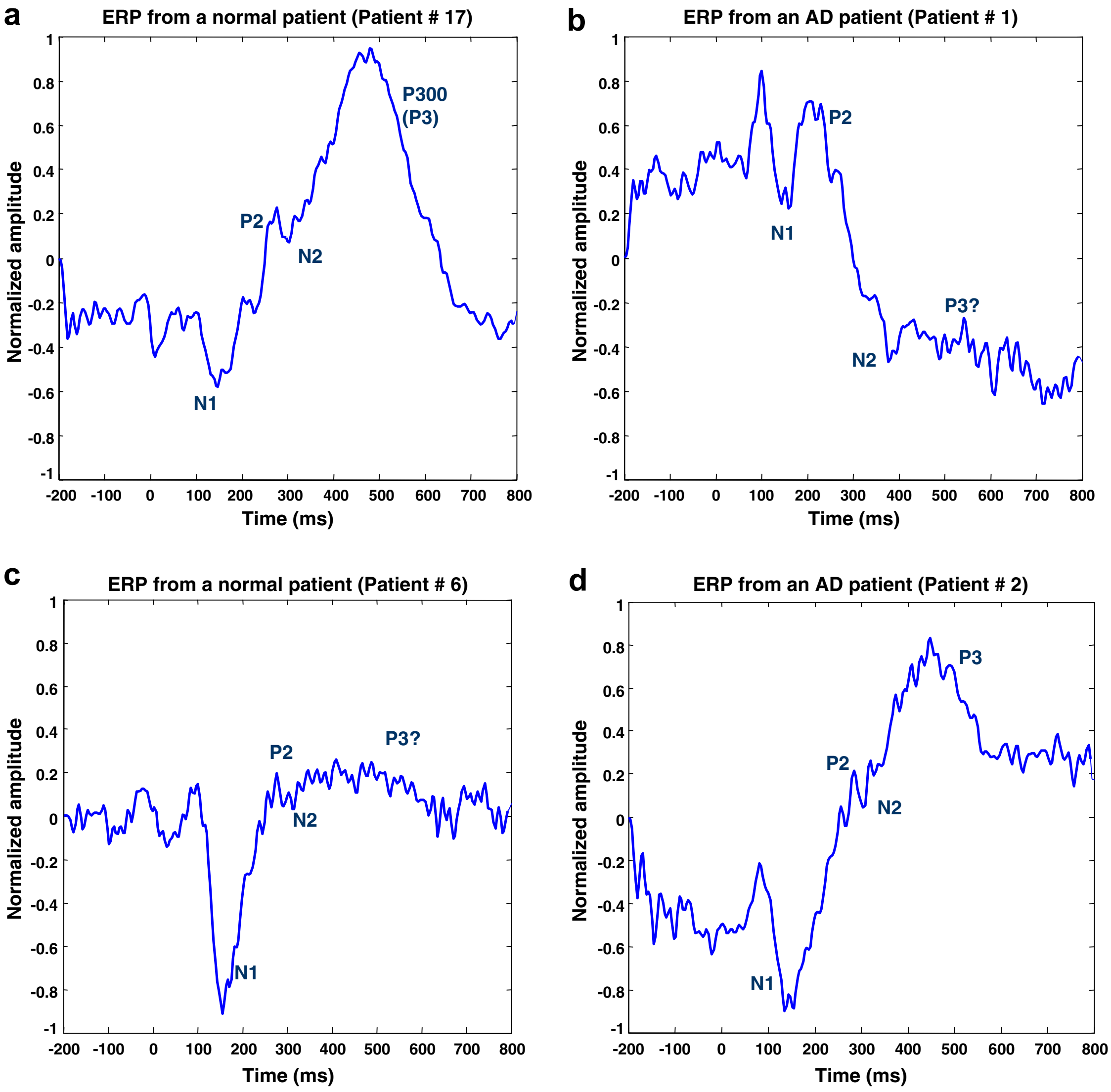


Fig. 1. (a) and (b) Expected P300 behavior from normal and AD patients; (c) and (d) not all individual cases follow this expected behavior.

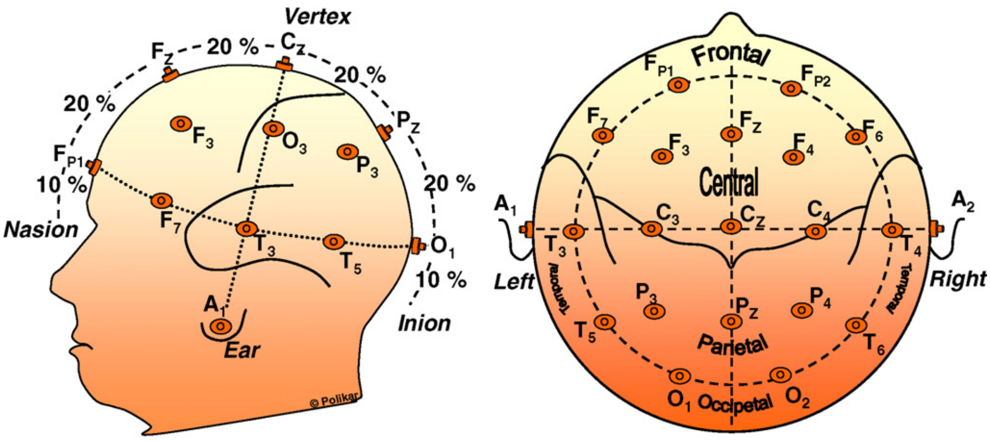


Fig. 2. The 10–20 International EEG electrode placement system.

AdaBoost algorithm, and borrows many of its algorithmic details; however, it has its differences from AdaBoost, described below, which are specifically designed for learning novel information from additional data.

1.2. Ensemble approaches and data fusion

In many applications that call for automated decision making, it is not unusual to receive data obtained from different sources that may provide complementary information. A suitable combination of such information is usually referred to as *data* or *information fusion*, and can lead to improved accuracy and confidence of the classification decision compared to a decision based on any of the individual data sources alone. Ensemble systems are naturally suited for such data fusion problems.

An ensemble based system, also known as a multiple classifier system (MCS), combines several, preferably diverse, classifiers. The diversity is typically achieved by using a different training dataset for each classifier, which then allows each classifier to generate different decision boundaries. The expectation is that each classifier will make a different error, and strategically combining these classifiers can reduce the total error. Since its humble beginnings with such seminal works including, but not limited to [23–29], research in multiple classifier systems has expanded rapidly, and has since become an important research topic [30]. Ensemble systems have appeared in literature under many creative names, such as composite classifier systems [23], stacked generalization [27], combination of multiple classifiers [31–33], dynamic classifier selection [33], classifier fusion [34,35], mixture of experts [36], committees of neural networks [37], or just classifier ensembles [30], among others. These approaches usually differ from each other in terms of the procedure by which individual classifiers are generated, and/or the procedure by which the classifiers are combined.

Most classifier combination approaches usually fall into one of two categories: classifier selection and classifier fusion [33,38]. In *classifier selection*, each classifier is trained to become an expert in some local area of the entire feature space. Given a data instance, the classifier trained with data closest to the vicinity of this instance is given the highest credit. In *classifier fusion*– not to be confused with data fusion – all classifiers are trained over the entire feature space. The classifier combination process then merges individual classifiers to obtain a single expert of superior performance, such as in bagging [39] or boosting based approaches [40,41]. The conditions under which – either or a combination of – classifier selection or classifier fusion may prove to be most useful are discussed in [42] .

Several combination rules are available, such as voting, sum, product or other combinations of posterior probabilities [28,29,35,43], fuzzy integral [44], Dempster–Shafer based combination [32,45], and more recently, decision templates [42,46]. Their comparison and theoretical analyses can be found in [29,35,47–49]. A sample of the immense literature on ensemble systems can be found in [30,50], and references therein.

We must mention that the word *''fusion''* that appears often in above-mentioned references usually refers to *''combination''* of classifiers with the goal of improving classifier generalization performance, by combining different pieces of information obtained from the same data source, and not necessarily for combining information coming from different data sources. Traditional methods for *data fusion* in this sense, originally developed for military applications such as target detection and tracking [51], are generally based on probability (Bayes theory, Kalman filtering, etc.) [52–54], evidence theory (Dempster–Shafer (DS) theory [55,56] and its variations [57,58]) , fuzzy and neural networks [44,59], or evolutionary algorithms [60].

Using the ensemble approach for data fusion applications, i.e., combining complementary knowledge from different data sources, while addressed in some studies [28,31,43], has in general been less explored – particularly for data with heterogeneous features. In this study, we therefore use a classifier-fusion type ensemble approach, not just for improving performance on a classification problem, but specifically for combining information from different sources – namely different channels of EEG and different frequency bands.

The rest of this paper is organized as follows: In Section 2, we describe the Learn^{++} algorithm in detail, adopted appropriately for data fusion applications, and provide guidance on specific implementation issues. In Section 3, we present the experimental setup for data collection. In Section 4, we present and interpret results, followed by conclusions and discussions in Section 5.

2. Learn^{++} for data fusion

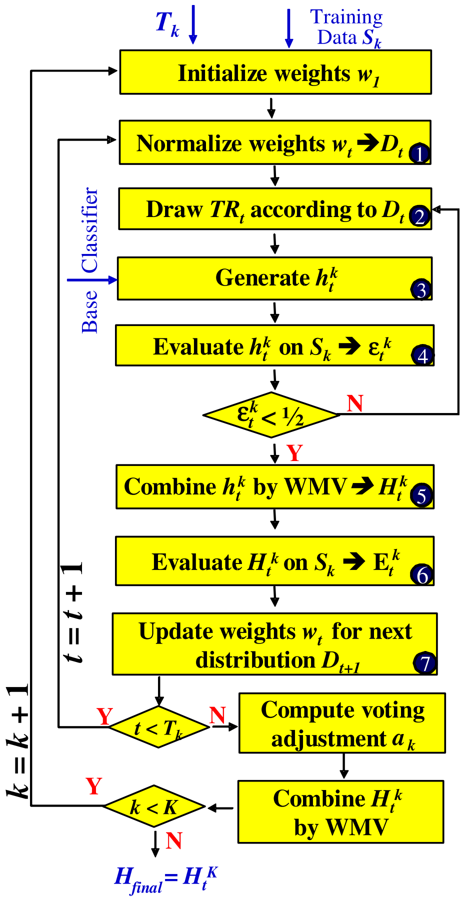
Learn^{++} was originally developed for incremental learning of novel information from new data – including from new classes – without forgetting the previously acquired knowledge, and without requiring access to previous data [61–64]. As in AdaBoost [41], Learn^{++} also generates an ensemble of classifiers, where each classifier is trained on a strategically updated distribution of the training data. Unlike AdaBoost, whose goal is to improve the performance of a classifier on a given dataset, Learn^{++} specifically targets learning from additional data: it generates an ensemble for each dataset that becomes available, and combines these ensembles to create an ensemble of ensembles, or a meta-ensemble of classifiers. More specifically, the distribution update rule through which consecutive classifiers are generated is different in Learn^{++}, and is geared towards learning the novel and discriminating information provided by each dataset that has not yet been learned by the *current ensemble*. Unlike AdaBoost, which updates its distribution based on the decision of the previously generated single classifier, Learn^{++} ties its distribution update directly to the ensemble decision. As we discuss the details below, the overall approach is then to generate an ensemble of classifiers for each dataset obtained from a different source, and appropriately combine the classifiers to extract additional information from subsequent data sources.

In the context of data fusion, we have K sources, each introducing a new dataset $\boldsymbol{DS}_k$, $k = 1, 2, \ldots, K$. For each dataset $\boldsymbol{DS}_k$ submitted to Learn^{++}, the algorithm inputs are (i) the training data $\boldsymbol{S}_k$ of m_k instances $\mathbf{x}_i$ along with their correct labels $y_i \in \Omega = \{\omega_1, \ldots, \omega_C\}$, $i = 1, 2, \ldots, m_k$, for C number of classes; (ii) a supervised classification algorithm *BaseClassifier*, generating individual classifiers (henceforth, hypotheses); and (iii) an integer T_k, the number of classifiers to be generated for the kth dataset. The pseudocode of the algorithm and its block diagram are provided in Figs. 3 and 4, respectively, and described below in detail.

The BaseClassifier can be any supervised classifier, whose *weakness* can be adjusted to ensure adequate diversity. This weakness can be controlled by adjusting training parameters (such as the size or error goal of a neural network) with respect to the complexity of the problem. However, a meaningful minimum performance is enforced: the probability of any classifier to produce the correct labels on a given training dataset, weighted proportionally to individual instances' probability of appearance, must be at least 1/2. If classifier outputs are class-conditionally independent, then the overall error monotonically decreases as new classifiers are added. Originally known as the Condorcet Jury Theorem (1786) [65,66], this condition is necessary and sufficient for a two-class problem ($C = 2$); and it is sufficient, but not necessary, for $C > 2$.

```
Input: For each feature set $\boldsymbol{DS}_k$, $k=1,2,\ldots,K$
  • Training data $S_k = [(\mathbf{x_i}, y_i)]$, i=1,…,$m_k$ with labels
    $y_i \in \Omega$, $\Omega = \{\omega_1, \cdots, \omega_C\}$.
  • Supervised algorithm BaseClassifier.
  • Integer $T_k$, specifying the number of classifiers.
Do for each $k=1,2,\ldots,K$:
  Initialize $w_1(i) = \boldsymbol{D}_1(i) = 1/m_k,\ \forall i,\ i = 1, 2, \cdots, m_k$
  Do for $t = 1,2,\ldots,T_k$:
    1. Set $\boldsymbol{D_t} = \boldsymbol{w_t} \Big/ \sum_{i=1}^{m_k} w_t(i)$ so that $\boldsymbol{D}_t$ is a distribution.
    2. Draw training $\boldsymbol{TR}_t$ subset from $\boldsymbol{D}_t$.
    3. Obtain $h_t^k$ by training with data $\boldsymbol{TR}_t$
    4. Calculate the error of h$_t$ $\varepsilon_t^k = \sum_{i:h_t^k(\mathbf{x}_i) \neq y_i} \boldsymbol{D_t}(i)$ on $S_k$
       If $\varepsilon_t^k > ½$, discard $h_t^k$ and →Step 2.
    5. Set $\beta_t^k = \varepsilon_t^k / (1 - \varepsilon_t^k)$. Obtain the composite hypothesis
       through weighted majority voting
         $H_t^k = \arg\max_{y \in \Omega} \sum_{t:h_t^k(\mathbf{x})=y} \log\left(1/\beta_t^k\right)$
    6. Compute error of $H_t^k$: $E_t^k = \sum_{i:H_t^k(\mathbf{x}_i) \neq y_i} \boldsymbol{D_t}(i)$
    7. Set $B_t^k = E_t^k / (1 - E_t^k)$, and update the weights:
       $w_{t+1}(i) = w_t(i) \times \begin{cases} B_t^k, & \mathit{if}\ \ H_t^k(\mathbf{x}_i) = y_i \\ 1, & \mathit{otherwise} \end{cases}$
  Compute voting weights adjustment factor
    $\alpha_k = \left( \sum_{i=1}^{m_k} \left[\left| H_{T_k}^k(\mathbf{x}_i) \neq y_i \right|\right] \right) \Big/ m_k$
Output the final hypothesis:
  $H_{final}(\mathbf{x}) = \arg\max_{y \in \Omega} \sum_{k=1}^{K} \frac{1}{\alpha_\mathrm{k}} \left( \sum_{t:h_t^k(\mathbf{x})=y} \log\left(\frac{1}{\beta_t^k}\right) \right)$
```

Fig. 3. Learn^{++} pseudocode for data fusion.



T_k
Training Data S_k
Initialize weights w_1
Normalize weights w_t → D_t
1
Draw TR_t according to D_t
2
Base Classifier
Generate h_t^k
3
Evaluate h_t^k on S_k → ε_t^k
4
$\varepsilon_t^k < ½$
N
Y
Combine h_t^k by WMV → H_t^k
5
Evaluate H_t^k on S_k → E_t^k
6
Update weights w_t for next distribution D_{t+1}
7
$t = t + 1$
Y
$t < T_k$
N
Compute voting adjustment a_k
$k = k + 1$
Y
$k < K$
Combine H_t^k by WMV
N
$H_{final} = H_t^K$

Fig. 4. Learn^{++} block diagram.

An iterative process sequentially generates each classifier of the ensemble: during the tth iteration, Learn^{++} trains the BaseClassifier on a judiciously selected subset $\boldsymbol{TR}_t$ of the current training data to generate hypothesis h_t^k. The training subset $\boldsymbol{TR}_t$ is drawn from the training data according to a distribution $\boldsymbol{D}_t$, which is obtained by normalizing a set of weights $\boldsymbol{w}_t$ maintained on the entire training data $\boldsymbol{S}_k$. The distribution $\boldsymbol{D}_t$ determines which instances of the training data are more likely to be selected into the training subset $\boldsymbol{TR}_t$. Unless a priori information indicates otherwise, this distribution is initially set to be uniform, by initializing $w_1(i) = 1/m_k$ $\forall i = 1, \ldots, m_k$, giving equal probability to each instance to be selected into $\boldsymbol{TR}_1$. At each subsequent iteration loop t, the weights previously adjusted at iteration $t - 1$ are normalized (in step 1 of the inner loop in Figs. 3 and 4)

$$\boldsymbol{D}_t = \boldsymbol{w}_t \Bigg/ \sum_{i=1}^{m_k} w_t(i) \tag{1}$$

to ensure a proper distribution. Training subset $\boldsymbol{TR}_t$ is drawn according to $\boldsymbol{D}_t$ (step 2), and the BaseClassifier is trained on $\boldsymbol{TR}_t$ (step 3). A hypothesis h_t^k is generated by the tth classifier, whose error ε_t^k is computed on the current dataset $\boldsymbol{S}_k$ as the sum of the distribution weights of the misclassified instances (step 4)

$$\varepsilon_t^k = \sum_{i:h_t^k(\mathbf{x}_i)\neq y_i} \boldsymbol{D}_t(i) = \sum_{i=1}^{m_k} \boldsymbol{D}_t(i)[|h_t^k(\mathbf{x}_i) \neq y_i|] \tag{2}$$

where $[|\cdot|]$ evaluates to 1, if the predicate holds true, and 0 otherwise. As mentioned above, we insist that this error be less than 1/2. If this is the case, the hypothesis h_t^k is accepted, and its error is normalized to obtain

$$\beta_t^k = \frac{\varepsilon_t^k}{1-\varepsilon_t^k}, \quad 0 < \beta_t^k < 1 \tag{3}$$

If $\varepsilon_t^k > 1/2$, the current hypothesis is discarded, and a new training subset is selected by returning to step 2. All hypotheses generated thus far are then combined using weighted majority voting [67], to obtain the *composite hypothesis* H_t^k (step 5), for which each hypothesis h_t^k is assigned a weight inversely proportional to its normalized error: those hypotheses with smaller training error are awarded a higher voting weight and thus have more say in the final classification decision. H_t^k then represents the current ensemble decision:

$$H_t^k = \arg\max_{y\in\Omega} \sum_{t:h_t^k(\mathbf{x})=y} \log(1/\beta_t^k) \tag{4}$$

It is relatively straightforward to prove that the weight selection of $\log(1/\beta_t^k)$ is optimum for weighted majority voting [30]. The error of the composite hypothesis H_t^k is then computed in a similar fashion as the sum of the distribution weights of the instances that are misclassified by the ensemble decision H_t^k (step 6)

$$E_t^k = \sum_{i:H_t^k(\mathbf{x}_i)\neq y_i} \boldsymbol{D}_t(i) = \sum_{i=1}^{m_k} \boldsymbol{D}_t(i)[|H_t^k(\mathbf{x}_i) \neq y_i|] \tag{5}$$

Since individual hypotheses that make up the composite hypothesis all have individual errors less than 1/2, so too will the composite error, i.e., $0 \leqslant E_t^k < 1/2$. The normalized composite error B_t^k can then be obtained as

$$B_t^k = \frac{E_t^k}{1-E_t^k}, \quad 0 < B_t^k < 1 \tag{6}$$

and is used for updating the distribution weights assigned to individual instances

$$w_{t+1}(i) = w_t(i) \times B_t^{k^{1-[|H_t^k(\mathbf{x}_i)\neq y_i|]}} = w_t(i) \times \begin{cases} B_t^k & \text{if } H_t^k(\mathbf{x}_i) = y_i \\ 1 & \text{otherwise} \end{cases} \tag{7}$$

Eq. (7) indicates that the distribution weights of the instances correctly classified by the composite hypothesis H_t^k are reduced by a factor of B_t^k. Effectively, this increases the weights of the misclassified instances making them more likely to be selected to the training subset of the next iteration. Readers familiar with AdaBoost have undoubtedly noticed the overall similarities, but also the key difference between the two algorithms: AdaBoost specifically targets improving the generalization performance of a weak learner on a single dataset by focusing on difficult instances that have been misclassified by the previous *hypothesis* h_t [41]. On the other hand, through the use of the composite hypothesis H_t, Learn^{++} specifically targets learning novel information from new data by focusing on those instances that are not yet learned by the existing ensemble. When Learn^{++} is acquiring novel information, the previously unseen or misclassified instances are precisely those not yet learned by the ensemble, forcing the algorithm to focus on instances carrying novel information. It can be argued that AdaBoost too looks (albeit indirectly) at the ensemble decision since, while based on a single hypothesis, the distribution update is cumulative. However, the update in Learn^{++} is directly tied to the ensemble decision, and hence been found to be more efficient in learning new information in our trials. The final hypothesis H_{final} is obtained by combining all hypotheses that have been generated thus far from all K data sources.

Fig. 5 conceptually illustrates the system level organization of the overall algorithm as structured for data fusion applications: an ensemble of classifiers is generated as described above for each of the feature sets, which are then combined through weighted majority voting. For data fusion applications, however, the performance based voting weights for each classifier, $\log(1/\beta_t^k)$, are further adjusted before final voting, based on expected or observed training performance on each data source: if prior information indicates that an individual data source is more reliable, a higher voting weight can be assigned to classifiers trained with such data. Alternatively, the weight adjustment can be based on the training performance of the ensemble on its own feature set. If such a strategy is chosen, the performance based weight of each classifier, $\log(1/\beta_t^k)$, is multiplied by the *reliability factor* of the feature set to which it belongs. This adjusted weight is then used to obtain the final hypothesis H_{final}:

$$H_{\text{final}}(\mathbf{x}) = \arg\max_{y\in\Omega} \sum_{k=1}^{K} \frac{1}{\alpha_k} \left(\sum_{t:h_t(\mathbf{x})=y} \log\left(\frac{1}{\beta_t^k}\right)\right) \tag{8}$$

where $1/\alpha_k$ is the reliability factor assigned to the ensemble trained on the kth feature set. In this work, α_k was chosen as the empirical error, that is, misclassification ratio of the final composite hypothesis on S_k:

$$\alpha_k = \left(\sum_{i=1}^{m_k} [|H_{T_k}^k(\mathbf{x}_i) \neq y_i|] \right) \Bigg/ m_k \tag{9}$$

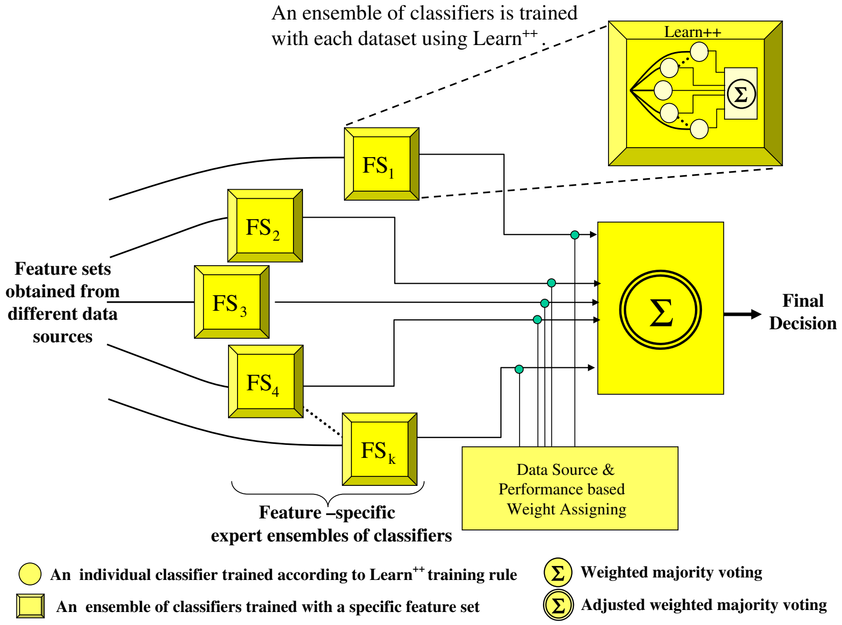

An ensemble of classifiers is trained with each dataset using Learn++
Learn++
FS1
FS2
FS3
FS4
FSk
Feature sets obtained from different data sources
Final Decision
Data Source & Performance based Weight Assigning
Feature –specific expert ensembles of classifiers
An individual classifier trained according to Learn++ training rule
An ensemble of classifiers trained with a specific feature set
Weighted majority voting
Adjusted weighted majority voting

Fig. 5. Schematic representation of the Learn^{++} based data fusion algorithm.

where H^k_{Tk} is the final composite hypothesis generated from the kth training data $\boldsymbol{S}_k$ of data source $\boldsymbol{DS}_k$.

Two implementation issues should be mentioned to prevent rare but pathological conditions causing deteriorated performance. First, classifiers with infinite voting-weight should be avoided. This issue arises when a classifier perfectly learns the entire training data (potentially over-fitting), resulting in $\beta^k_t = 0$, and hence the voting weight of h^k_t to be infinite. The outcome is then a despotic h^k_t with the sole power of decision making. This situation can be avoided either by making classifiers weaker (so that the training error exceeds zero), or by adding a small adjustment factor (0.01 usually works rather well) to β^k_t.

Second, unless there is prior information to choose otherwise, the number of classifiers generated for each dataset should be (at least, approximately) the same. The number of classifiers is usually selected such that the performance on a separate validation dataset is maximized, where classifiers are added to the ensemble until addition of classifiers no longer contributes to performance improvement. However, it is not unusual for the ensemble performance to stay constant, or only slightly fluctuate for a large number of classifiers. In such cases, a large number of classifiers may be retained despite the lack of meaningful performance gain. Apart from increased computational complexity and potential for over-fitting, unnecessarily large number of classifiers generated with any of the feature sets also causes a bias in the final classification towards the data source that has more classifiers. This situation can be avoided by applying regularization to the validation, or heuristically picking the number of classifiers to be the same for each data source.

Simulation results of Learn^{++} on incremental learning of several scenarios, as well as comparisons to the other similar methods can be found in [61,68,69], and its ability of confidence estimation in [63,70]. Results on data fusion of EEG data are presented below, following the experimental setup.

3. Experimental setup and feature extraction

3.1. Research subjects and the gold standard

Considering that the current best method of diagnosis is clinical evaluation through a neuropsychological test, the outcome of such a test constitutes the gold standard for this analysis. To date, 52 subjects have been recruited by the Memory Disorders Clinic of University of Pennsylvania, specifically for this study. The following inclusion and exclusion criteria were used for the probable AD and cognitively normal cohorts.

Inclusion criteria for cognitively normal cohort: (i) age $>$ 60; (ii) Clinical Dementia Rating score $=$ 0; (iii) Mini Mental State Exam score $>$ 26; (iv) no indication of functional or cognitive decline during the two years prior to enrollment based on a detailed interview with the subject's knowledgeable informant.

Exclusion criteria for cognitively normal cohort: (i) evidence of any central nervous system neurological disease (e.g., stroke, multiple sclerosis, Parkinson's disease, etc.) by history or exam; (ii) use of sedative, anxiolytic or antidepressant medications within 48 h of ERP acquisition.

Inclusion criteria for AD cohort: (i) age $>$ 60; (ii) Clinical Dementia Rating score $\geqslant$ 0.50; (iii) Mini Mental State Exam score $\leqslant$ 26; (iv) presence of functional and cognitive decline over the previous 12 months based on detailed interview with a knowledgeable informant; (v) satisfaction of NINCDS–ADRDA (National Institute of Neurological and Communicative Disorders and Stroke–Alzheimer's Disease and Related Disorders Association) criteria for probable AD [71].

Exclusion criteria for AD cohort: Same as for the cognitively normal controls.

All subjects received a thorough medical history and neurological exam. Key demographic and medical information, including current medications (prescription, over-the-counter, and complementary alternative medications) were noted. The evaluation included standardized assessments for overall impairment, cognitive impairment, functional impairment, extrapyramidal signs, behavioral changes and depression. The clinical diagnosis was made as a result of these analyses, as described by the NINCDS–ADRDA criteria for probable AD [71].

The inclusion criteria for AD cohort were designed to ensure that subjects were at the earliest clinical stage of the disease. One measure to evaluate the severity of the disease is the Mini Mental State Exam (MMSE), a widely used standardized exam designed to assess orientation, attention, immediate and short-term recall, language, and the ability to follow simple verbal and written commands. MMSE also provides a total score, on a scale of 0 to 30, that provides an indication of cognitive function. Cognitive performance shows an inverse relationship between MMSE scores and age/education, ranging from a median of 29 for those 18–24 years of age, to 25 for individuals 80 years of age and older. The median MMSE score is 29 for individuals with at least 9 years of schooling, 26 for those with 5–8 years of schooling, and 22 for those with 0–4 years of schooling [72,73].

We emphasize that MMSE alone is not used for diagnosis, but just as one measure for assessing the severity of disease. The AD diagnosis itself is made based on the above-mentioned NINCDS–ADRDA criteria for probable AD. Of all subjects who were diagnosed with probable AD, we included only those with an MMSE score of 24 or above (the number of years of schooling was also considered) to ensure to include only those at the earliest stages of the disease. Of the 52 subjects recruited to date, 28 of them were AD patients ($\mu_{\text{Age}} = 79$, $\mu_{\text{MMSE}} = 25$) and 24 were cognitively normal individuals ($\mu_{\text{Age}} = 76$, $\mu_{\text{MMSE}} = 29$).

3.2. Event related potentials (ERPs) acquisition protocol

The ERPs were obtained using an auditory oddball paradigm while the subjects were comfortably seated in a specially designated room. The protocol described by Yamaguchi et al. [4] was used with slight modifications. Binaural audiometric thresholds were first determined for each subject using a 1 kHz tone. The evoked response stimulus was presented to both ears using stereo earphones at 60 dB above each individual's auditory threshold. Each stimulus consisted of tone bursts 100 ms in duration, including 5 ms onset and offset envelops. A total of 1000 such stimuli of frequent 1 kHz normal tones (65%), infrequent 2 kHz oddball (target) tones (20%), and novel sounds (15%) were delivered to each subject with an inter-stimulus interval of 1.0–1.3 s. Novel sounds consisted of 60 unique digitally recorded environmental sounds that were edited to a 200 ms duration. To maintain the novelty of the stimuli, each novel sound was presented only once. The subjects were instructed to press a button each time they heard the 2 kHz oddball tone. With frequent breaks (3 min of rest every 5 min), data collection typically took less than 30 min. The experimental session was preceded by a 1-min practice session without the novel sounds.

ERPs were recorded from 19 electrodes embedded in an elastic cap. The electrode impedances were kept below 20 kΩ. Artifactual epochs were removed by the EEG technician. The potentials were then amplified, digitized at 256 Hz/channel, lowpass filtered, averaged (40–90 oddball/novel tones per patient), notched filtered at 59–61 Hz, and baselined with the pre-stimulus interval for a final 256-sample long signal.

3.3. Feature extraction

Multiresolution wavelet analysis determines time localizations of spectral components, providing a time–frequency representation of the signal being analyzed. Such an analysis is particularly well suited for non-stationary signals, such as the ERPs, whose spectral content varies in time. Among many time–frequency representations, the discrete wavelet transform (DWT) has become increasingly popular due its ability to solve a diverse set of problems, including data compression, biomedical signal analysis, feature extraction, noise suppression, density estimation, and function approximation – all with modest computational expense. DWT is a well established technique; for brevity yet completeness, only an overview is therefore provided here. Interested readers are referred to [74], and references within, for additional details.

The DWT analyzes the signal at different frequency bands with different resolutions using a decomposition process. The DWT utilizes two sets of functions, scaling and wavelet functions, each associated with lowpass and highpass filters, respectively. Decomposition of the signal into different frequency bands is accomplished by successive highpass and lowpass filtering of the time domain signal.

The original time domain signal $x(t)$ sampled at 256 samples/s creates the discrete time signal $x[n]$ which is passed through a halfband highpass filter $g[n]$ and a lowpass filter $h[n]$. In terms of angular frequency, the highest frequency in the original signal is π rad/s, corresponding to the linear frequency of 128 Hz. According to Nyquist's rule, half the samples can be removed after the filtering, since the bandwidth of the signal is now π/2 rad/s. Therefore every other sample in the signal can be discarded. This is one level of decomposition and can be expressed as follows:

$$y_{\text{high}}[k] = \sum_{n} x[n] \cdot g[2k - n] \tag{10}$$

$$y_{\text{low}}[k] = \sum_{n} x[n] \cdot h[2k - n] \tag{11}$$

where $y_{\text{high}}[k]$ and $y_{\text{low}}[k]$ are the outputs of the highpass and lowpass filters after the subsampling, and are referred to as *detail coefficients* and *approximation coefficients*, respectively. This procedure, known as subband coding, is repeated by decomposing the approximation coefficients until further decomposition is not possible (due to loss of samples through downsampling). The detail coefficients $\boldsymbol{d}_i$ at level i then constitute Level i DWT coefficients. At each level, the successive filtering and subsampling result in half the time resolution and double the frequency resolution. Therefore, each level of decomposition analyzes the signal at different frequency ranges and different resolutions, hence multiresolution analysis. Fig. 6 illustrates this procedure, where the bandwidth of the signal at every level is marked on the figure as "B".

Fig. 7 shows the eight signals obtained from 7-level decomposition of a sample ERP (from a cognitively normal patient). For 256-sample signal $x[n]$ and using Daubechies-4 wavelets (of length 8), these levels correspond to the following frequency bands: d_1: 64–128 Hz (132 coefficients); d_2: 32–64 Hz (69 coefficients); d_3: 16–32 Hz (38 coefficients); d_4: 8–16 Hz (22 coefficients); d_5: 4–8 Hz (14 coefficients); d_6: 2–4 Hz (10 coefficients); d_7: 1–2 Hz (8 coefficients); and a_7: 0–1 Hz (8 coefficients). The coefficient amplitudes are in general higher at lower frequencies, in line with where we would expect to see most of the ERP information. We should also note that as the signal length gets smaller at each level, the boundary effects of the filtering becomes more prominent. Therefore, only those coefficients corresponding approximately to 100–600 ms after the stimulus were used in the analysis. The actual number of coefficients used in our analyses are shown in Table 1.

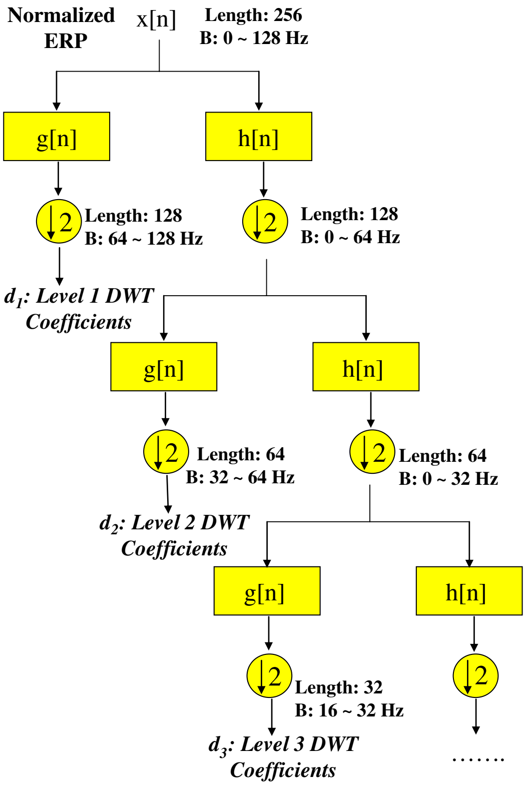


Fig. 6. DWT subband coding algorithm.

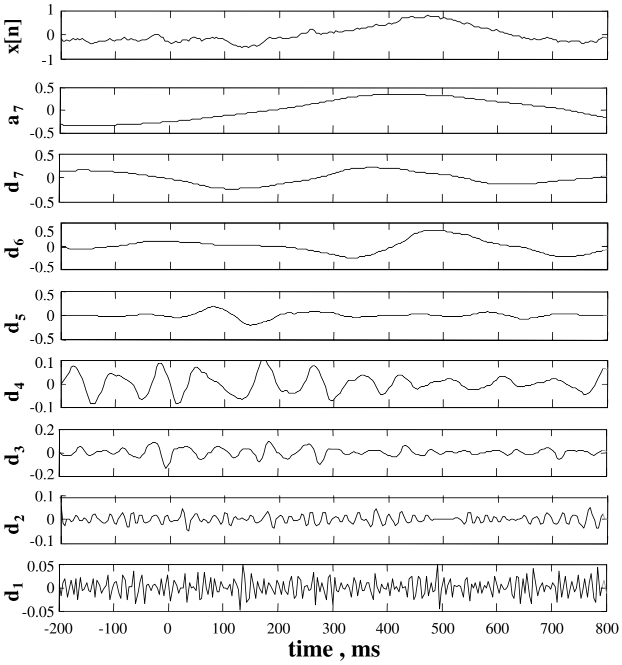


Fig. 7. Sample ERP decomposition.

4. Results

Previous studies mentioned in the introduction have looked at data primarily from the Pz electrode, or from Pz, Fz and Cz electrodes, one at a time, and primarily in response to target tones only. Therefore, our goal was to determine whether a combination of signals obtained from different electrodes, in response to different stimuli, and analyzed in different frequency bands provide a better diagnostic performance.

Considering that there are three different electrode locations, two stimulus tones, and eight frequency bands, the natural question is then "which electrode – stimulus tone – frequency band combination provides the most information?" To answer this question, we have individually analyzed datasets obtained from all 48 three-tuple combinations, choosing one electrode, one stimulus tone, and one frequency band. The five datasets that provided the best individual performances are shown in Table 1. Four of the five highest performing datasets corresponded to data in the 1–4 Hz range, where the P300 is known to reside, indicating that P300 is indeed influential in AD diagnosis. However, none of the individual data sources

Table 1
Five highest performing electrode/frequency band/stimulus type combinations

Electrode	Response to	Abbreviation	# of coefficients	Frequency band (Hz)	Performance (%)
Pz	Novel sounds	NPz_1	4	1–2	75.0
Pz	Novel sounds	NPz_2	5	2–4	63.2
Fz	Target sounds	TFz	6	4–8	63.6
Cz	Target sounds	TCz	5	2–4	63.8
Pz	Target sounds	TPz	5	2–4	60.4

provide a particularly stellar performance, except perhaps the Pz electrode with novel sounds at 1–2 Hz range (NPz_1).

Leave-one-out cross validation, widely considered to be the best estimate of the true generalization performance of a classifier on small datasets, was used in our experiments. All performance figures in Tables 1 and 2 are therefore obtained as averages of five independent leave-one-out trials using an ensemble of five classifiers: in each trial, a 5-classifier Learn^{++} ensemble was trained using 51 of the 52 patient data, and the ensemble was then evaluated on the remaining 52nd patient. The base classifier was a single hidden layer MLP with 10 hidden layer nodes, 2 output nodes (one for each class), and an error goal of 0.01. The number of input nodes was the number of DWT coefficients shown in Table 1. This process was repeated 52 times, in each case testing on a different patient. The average of these 52 *five-classifier ensembles* constitutes *one* leave-one-out trial. All performance figures are then averages of *five* such independent leave-one-out trials.

Note that no data fusion is yet applied. For each of the five datasets mentioned in Table 1, Learn^{++} is trained only on that single dataset ($k = 1$, and hence, the α_k parameter does not yet apply) to determine the individual ensemble performance that can be achieved by that dataset alone.

The individual ensembles were then fused using Learn^{++}, as modified for data fusion applications (Fig. 1). There are a total of 26 possible 2-, 3-, 4- or 5-way fusion for five datasets shown above. For brevity, we report the top five data-fusion performances. In Table 2, performances of individual datasets (from Table 1) are given first, followed by the data fusion performance obtained by Learn^{++} combination of individual ensembles. Since each ensemble had five classifiers, a 2-way combination has a total of 10 classifiers (and a 4-way fusion has 20). Note that data fusion performances are also averages of five independent leave-one-out trials, but now each trial itself is an average of 52 ten-classifier ensembles – if two datasets are fused, or 52 twenty-classifier ensembles – if four datasets are fused.

Sensitivity and specificity numbers are also provided, which are commonly used in medical diagnostics. In medical terminology, sensitivity is the probability that a symptom is present (test is positive, or the ensemble declares AD) given that the person has the disease (the true positive). The sensitivity measures the ability of the test (the classifier) in identifying those who have the disease. The specificity, however, measures the ability of the test in identifying those who do not have the disease. Hence, specificity is the probability that a symptom is not present (test is negative, or the ensemble declares the patient as normal) given that the person does not have the disease (true negative).

Table 2 indicates that the diagnostic performance of the fusion of any 2- or 4-way combinations is better than the performance of the any of the individual data sources. Combinations including NPz_1 performed better than the others, as expected, since the NPz_1 dataset provided the best single performance. What is interesting, however, is that NPz_1 and TFz combination provided significantly better data fusion performance than the NPz_1 and NPz_2 combination. This indicates that – given the information provided by NPz_1 – there is more complementary information in TFz data than in the NPz_2 data. On the other hand, the 4-way combination of the four best performing datasets did not perform better than the NPz_1 and TFz combination, indicating that there is no additional complementary information provided by the remaining two datasets (TCz and NPz_2) beyond what is already provided in the NPz_1 and TFz combination. Furthermore, combining data obtained in response to novel *and* target tones appears to perform better than combining data from target *or* novel tones only, indicating that target and novel tones may provide complementary information – but only if recorded at different electrodes. The TFz & NFz and TCz & NCz combinations, for example, only had mid-60% range performance (not shown in Table 2).

The NPz_1 and TFz fusion performance of 79.2%, best data fusion performance achieved so far, may not appear

Table 2
Data fusion performances of various combinations of datasets

	Fused datasets											
	TFz + TCz		TCz + TPz		$NPz_1 + NPz_2$		NPz_1 + TFz		TCz + TFz + $NPz_1 + NPz_2$			
Individual performances	63.6	63.8	63.8	60.4	75.0	63.2	75.0	63.6	63.8	63.6	75.0	63.2
Fusion performance (%)	68.8		70.0		75.4		79.2		78.8			
Sensitivity (%)	72.2		71.4		63.6		74.3		77.1			
Specificity (%)	65.2		68.3		89.4		85.0		80.8			

as particularly impressive, as the improvement is only about 5% over the performance of the best single dataset of NPz_1. However 79.2% significantly exceeds that of other HMO based community clinics, and is clinically considered to be very significant, considering the difficulty of identifying AD at its earliest stages. We should add that the best of the five leave-one-out trials (whose average was 79.2%) was 80.8%. We include this information, since most studies typically report a single leave-one-out performance.

An interesting observation can also be made from the sensitivity and specificity figures. On average, sensitivity is higher than specificity for ensembles trained with data in response to target tones (TFz&TPz and TFz&TCz), whereas the opposite is true for the ensemble trained with data in response to novel tones (NPz_1 and NPz_2). This indicates that the target tone provides better information in identifying AD patients. This is not surprising, as it is these target tones to which AD patients have difficulty responding. However, the results also indicate that the novel tones provide better information in identifying normal patients. Since novel tones have only recently been used, this piece of information – while logical and expected – is also extremely satisfying. Hence the fusion of novel NPz_1 and target TFz data, apart from giving the best overall performance, also provide a more balanced sensitivity and specificity combination.

5. Discussion and conclusions

We have evaluated the Learn^{++} algorithm – originally developed for incremental learning and adapted for data fusion – in addressing a challenging real world data fusion problem. Three specific characteristics of Learn^{++} makes it a particularly good match for data fusion applications, such as the one featured in this paper, where data from different sources need to be combined: (i) the ensemble structure provides a natural mechanism to combine heterogeneous features, (ii) the sequential generation of classifiers based on the ensemble performance allows efficient learning of complementary information in each dataset; and (iii) weighted majority voting with integrated reliability factor allows giving a higher weight to those ensembles trained on more reliable/informative data sources.

The application presented in this work seeks the diagnostic identification of AD vs. normal patients based on their ERP recordings. Of particular interest – which makes the problem particularly challenging – is diagnosis of the disease at its earliest possible stages. Based on the results presented above, we draw the following conclusions: (i) using wavelet analysis to extract features of the ERPs, followed by ensemble based data fusion appears to be an effective tool for early diagnosis of AD; (ii) the fusion of ensembles trained on individual data at different frequency bands that are obtained from different electrodes, typically perform better than a similarly configured ensemble trained on any of the individual datasets. This demonstrates that the algorithm can extract complementary information from different sources, if such information exists; (iii) the fusion approach provides insight into which electrode/frequency interval/stimulus type combination provides the most information, and hence can be used as a feature selection procedure. Furthermore, the sensitivity – specificity analysis provides particular insight to the most effective use of target vs. novel tones; (iv) the approach is non-invasive, cost-effective, and can be made readily available to community clinics, since EEG recording technology is well established and widely available; (v) having tried several BaseClassifier architectures and error goals, the algorithms seems to be quite invariant to minor changes in these parameters. Therefore the approach is expected to be a stable and effective one; (vi) finally, the approach seems to meet or exceed the current performances of community based clinical evaluations, even at detecting the disease at its earliest stage. This is clinically significant, as the proposed approach – when fully developed – can provide an initial screening for majority of the patients at an early stage. If the patient is in fact normal, cost savings can be very significant, as they may not need the costly clinical evaluation. If, on the other hand, the approach indicates AD, this would provide an early warning for a need for clinical evaluation. An early diagnosis, resulting in early intervention and appropriate medication, can then add many years to the patient's life, not to mention, significantly improving quality of life for the patient as well as their caregivers.

Our future work includes repeating virtually all experiments as additional patients are recruited (up to 80 will be recruited). We will also be expanding this analysis to include a third cohort: patients suffering from Parkinson's disease, about 30% of whom eventually develop dementia. Formal analysis of the algorithm on several different scenarios of data fusion and additional EEG channels will also be part of our future efforts.

Acknowledgements

This work is supported by National Institute on Aging/National Institutes of Health under grant number P30 AG10124–R01 AG022272, and by National Science Foundation under Grant No. ECS-0239090. The anonymous reviewers' comments have been most informative and helpful in improving this paper. They are greatly appreciated.

References

[1] Alzheimer's Disease International, About Alzheimer's Disease. <http://www.alz.co.uk/alzheimers/faq.html#howmany> (accessed 21.10.06).

[2] Alzheimer's Association, Basic Facts and Statistics. <http://www.alz.org/AboutAD/statistics.asp> (accessed 21.10.06).

[3] A. Lim, W. Kukull, D. Nochlin, J. Leverenz, W. McCormick, J. Bowen, L. Teri, J. Thompson, E. Peskind, M. Raskind, E. Larson, Clinico-neuropathological correlation of Alzheimer's disease in a

community-based case series, Journal of the American Geriatrics Society 47 (1999) 564–569.

[4] S. Yamaguchi, H. Tsuchiya, S. Yamagata, G. Toyoda, S. Kobayashi, Event-related brain potentials in response to novel sounds in dementia, Clinical Neurophysiology 111 (2000) 195–203.

[5] J. Polich, P300 and Alzheimer's disease, Biomedicine and Pharmacotherapy 43 (1989) 493–499.

[6] J. Polich, C. Ladish, F.E. Bloom, P300 assessment of early Alzheimer's disease, Electroencephalography and Clinical Neurophysiology 77 (1990) 179–189.

[7] S. Yamaguchi, H. Tsuchiya, S. Yamagata, G. Toyoda, S. Kobayashi, Event-related brain potentials in response to novel sounds in dementia, Clinical Neurophysiology 111 (2000) 195–203.

[8] A. Ademoglu, T. Demiralp, J. Yordanova, V. Kolev, M. Devrim, Decomposition of event-related brain potentials into multicomponents using wavelet transform, Applied Signal Processing 5 (1998) 142–151.

[9] S. Aviyente, L.A.W. Brakel, R.K. Kushwaha, M. Snodgrass, H. Shevrin, W.J. Williams, Characterization of event related potentials using information theoretic distance measures, IEEE Transactions on Biomedical Engineering 51 (2004) 737–743.

[10] E. Basar, M. Schurmann, T. Demiralp, C. Basar-Eroglu, A. Ademoglu, Event-related oscillations are 'real brain responses' – wavelet analysis and new strategies, International Journal of Psychophysiology 39 (2001) 91–127.

[11] T. Demiralp, A. Ademoglu, Y. Istefanopulos, C. Basar-Eroglu, E. Basar, Wavelet analysis of oddball P300, International Journal of Psychophysiology 39 (2001) 221–227.

[12] T. Demiralp, A. Ademoglu, Decomposition of event-related brain potentials into multiple functional components using wavelet transform, Clinical Electroencephalography 32 (2001) 122–138.

[13] R.M. Chapman, G.H. Nowlis, J.W. McCrary, J.A. Chapman, T.C. Sandoval, M.D. Guillily, M.N. Gardner, L.A. Reilly, Brain event-related potentials: diagnosing early-stage Alzheimer's disease, Neurobiology of Aging, in press.

[14] S.Y. Cho, B.Y. Kim, E.H. Park, J.W. Kim, W.W. Whang, S.K. Han, H.Y. Kim, Automatic recognition of Alzheimer's disease with single channel EEG recording, International Conference of the IEEE Engineering in Medicine and Biology 3 (2003) 2655–2658.

[15] A.A. Petrosian, D.V. Prokhorov, W. Lajara-Nanson, R.B. Schiffer, Recurrent neural network-based approach for early recognition of Alzheimer's disease in EEG, Clinical Neurophysiology 112 (2001) 1378–1387.

[16] S. Yagneswaran, M. Baker, A. Petrosian, Power frequency and wavelet characteristics in differentiating between normal and Alzheimer EEG, in: IEEE Engineering in Medicine and Biology 24th Annual Conference, vol. 1, 2002, pp. 46–47.

[17] G. Jacques, J. Frymiare, C. Kounios, C. Clark, R. Polikar, Multiresolution analysis for early diagnosis of Alzheimer's disease, in: 26th Annual International Conference of IEEE Engineering in Medicine and Biology Society, San Francisco, CA, 2004, pp. 251–254.

[18] R. Polikar, F. Keinert, Wavelet analysis of event related potentials for early diagnosis of Alzheimer's disease, in: A. Petrosian, F.G. Meyer (Eds.), Wavelets in Signal and Image Analysism, From Theory to Practice, Kluwer Academic Publishers, Boston, 2001.

[19] J. Wallerius, L.J. Trejo, R. Matthews, R. Rosipal, J.A. Caldwell, Robust feature extraction and classification of EEG spectra for real-time classification of cognitive state, 11th International Conference on Human Computer Interaction, Las Vegas, Nevada, 2005.

[20] L.J. Trejo, R. Rosipal, B. Matthews, Brain–computer interfaces for 1-D and 2-D cursor control: designs using volitional control of the EEG spectrum or steady-state visual evoked potentials, IEEE Transactions on Neural Systems and Rehabilitation Engineering 14 (2006) 225–229.

[21] Y.O. Halchenko, S.J. Hanson, B.A. Pearlmutter, Multimodel integration: fMRI, MRI, EEG, MEG, in: L. Landini, V. Positano, M.F. Santarelli (Eds.), Advanced Image Processing in Magnetic Resonance Imaging, CRC Press, 2005.

[22] B.H. Jansen, A. Allam, P. Kota, K. Lachance, A. Osho, K. Sundaresan, An exploratory study of factors affecting single trial P300 detection, IEEE Transactions on Biomedical Engineering 51 (2004) 975–978.

[23] B.V. Dasarathy, B.V. Sheela, Composite classifier system design: concepts and methodology, Proceedings of the IEEE 67 (1979) 708–713.

[24] L.K. Hansen, P. Salamon, Neural network ensembles, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1990) 993–1001.

[25] R.E. Schapire, The strength of weak learnability, Machine Learning 5 (1990) 197–227.

[26] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton, Adaptive mixtures of local experts, Neural Computation 3 (1991) 79–87.

[27] D.H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259.

[28] T.K. Ho, J.J. Hull, S.N. Srihari, Decision combination in multiple classifier systems, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (1994) 66–75.

[29] J. Kittler, M. Hatef, R.P.W. Duin, J. Mates, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 226–239.

[30] L.I. Kuncheva, Combining Pattern Classifiers, Methods and Algorithms, Wiley Interscience, Hobokem, NJ, 2005.

[31] L. Xu, A. Krzyzak, C.Y. Suen, Methods of combining multiple classifiers and their applications to handwriting recognition, IEEE Transactions on Systems, Man and Cybernetics 22 (1992) 418–435.

[32] G. Rogova, Combining the results of several neural network classifiers, Neural Networks 7 (1994) 777–781.

[33] K. Woods, W.P.J. Kegelmeyer, K. Bowyer, Combination of multiple classifiers using local accuracy estimates, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 405–410.

[34] L.I. Kuncheva, J.C. Bezdek, R.P.W. Duin, Decision templates for multiple classifier fusion: an experimental comparison, Pattern Recognition 34 (2001) 299–314.

[35] L.I. Kuncheva, A theoretical study on six classifier fusion strategies, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 281–286.

[36] M.J. Jordan, R.A. Jacobs, Hierarchical mixtures of experts and the EM algorithm, Neural Computation 6 (1994) 181–214.

[37] H. Drucker, C. Cortes, L.D. Jackel, Y. LeCun, V. Vapnik, Boosting and other ensemble methods, Neural Computation 6 (1994) 1289–1301.

[38] L.I. Kuncheva, J.C. Bezdek, R.P.W. Duin, Decision templates for multiple classifier fusion: an experimental comparison, Pattern Recognition 34 (2001) 299–314.

[39] L. Breiman, Bagging predictors, Machine Learning 24 (1996) 123–140.

[40] R.E. Schapire, The strength of weak learnability, Machine Learning 5 (1990) 197–227.

[41] Y. Freund, R.E. Schapire, Decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1997) 119–139.

[42] L.I. Kuncheva, Switching between selection and fusion in combining classifiers: an experiment, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 32 (2002) 146–156.

[43] D.M.J. Tax, M. van Breukelen, R.P.W. Duin, J. Kittler, Combining multiple classifiers by averaging or by multiplying? Pattern Recognition 33 (2000) 1475–1485.

[44] S.B. Cho, J.H. Kim, Multiple network fusion using fuzzy logic, IEEE Transactions on Neural Networks 6 (1995) 497–501.

[45] Y. Lu, Knowledge integration in a multiple classifier system, Applied Intelligence 6 (1996) 75–86.

[46] L.I. Kuncheva, J.C. Bezdek, R.P.W. Duin, Decision templates for multiple classifier fusion: an experimental comparison, Pattern Recognition 34 (2001) 299–314.

[47] K. Tumer, J. Ghosh, Analysis of decision boundaries in linearly combined neural classifiers, Pattern Recognition 29 (1996) 341–348.

[48] N.S.V. Rao, On fusers that perform better than best sensor, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 904–909.

[49] G. Fumera, F. Roli, A theoretical and experimental analysis of linear combiners for multiple classifier systems, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 942–956.

[50] Various Authors, Proceedings of International Workshop on Multiple Classifier Systems (2000–2005), 2005.

[51] D.L.Ed. Hall, Handbook of Multisensor Data Fusion, CRC Press, 2001.

[52] D.M. Buede, P. Girardi, A target identification comparison of Bayesian and Dempster–Shafer multisensor fusion, IEEE Transactions on Systems, Man and Cybernetics (A) 27 (1997) 569–577.

[53] M.B. Hurley, An extension of statistical decision theory with information theoretic cost functions to decision fusion: Part II, Information Fusion 6 (2005) 165–174.

[54] J.Z. Sasiadek, Sensor fusion, Annual Reviews in Control 26 (2002) 203–228.

[55] A.P. Dempster, Upper and lower probabilities induced by multivalued mappings, Annals of Mathematical Statistics 38 (1967) 325–339.

[56] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, NJ, 1976.

[57] D. Fixsen, R.P.S. Mahler, The modified Dempster–Shafer approach to classification, IEEE Transactions on Systems, Man and Cybernetics (A) 27 (1997) 96–104.

[58] R.R. Murphy, Dempster–Shafer theory for sensor fusion in autonomous mobile robots, IEEE Transactions on Robotics and Automation 14 (1998) 197–206.

[59] G.A. Carpenter, S. Martens, O.J. Ogas, Self-organizing information fusion and hierarchical knowledge discovery: a new framework using ARTMAP neural networks, Neural Networks 18 (2005) 287–295.

[60] I.V. Maslov, I. Gertner, Multi-sensor fusion: an evolutionary algorithm approach, Information Fusion 7 (2006) 304–330.

[61] R. Polikar, L. Udpa, S.S. Udpa, V. Honavar, Learn^{++}: an incremental learning algorithm for supervised neural networks, IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews 31 (2001) 497–508.

[62] R. Polikar, J. Byorick, S. Krause, A. Marino, M. Moreton, Learn^{++}: a classifier independent incremental learning algorithm for supervised neural networks, in: International Joint Conference on Neural Networks, Honolulu, HI, vol. 2, 2002, 1742–1747.

[63] R. Polikar, L. Udpa, S. Udpa, V. Honavar, An incremental learning algorithm with confidence estimation for automated identification of NDE signals, IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control 51 (2004) 990–1001.

[64] M. Muhlbaier, A. Topalis, R. Polikar, Learn^{++}.MT: a new approach to incremental learning, in: F. Roli, J. Kittler, T. Windeatt (Eds.), 5th International Workshop on Multiple Classifiers Systems, Lecture Notes in Computer Science, Cagliari, Italy, vol. 3077, 2004, pp. 52–61.

[65] P.J. Boland, Majority system and the Condorcet Jury theorem, Statistician 38 (1989) 181–189.

[66] D. Berend, J. Paroush, When is Condorcet's Jury Theorem valid? Social Choice and Welfare 15 (1998) 481–488.

[67] N. Littlestone, M. Warmuth, Weighted majority algorithm, Information and Computation 108 (1994) 212–261.

[68] M. Muhlbaier, A. Topalis, R. Polikar, Learn^{++}.MT: a new approach to incremental learning, in: F. Roli, J. Kittler, T. Windeatt (Eds.), 5th International Workshop on Multiple Classifiers Systems, Lecture Notes in Computer Science, Cagliari, Italy, vol. 3077, 2004, pp. 52–61.

[69] A. Gangardiwala, R. Polikar, Dynamically weighted majority voting for incremental learning and comparison of three boosting based approaches, IEEE International Joint Conference on Neural Networks, Montreal, QB, vol. 2, 2005, pp. 1131–1136.

[70] M. Muhlbaier, A. Topalis, R. Polikar, Ensemble confidence estimates posterior probability, in: N.C. Oza, R. Polikar, J. Kittler, F. Roli (Eds.), 6th International Workshop on Multiple Classifier Systems, Monterey, CA, vol. 3541, 2005, pp. 326–335.

[71] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, 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 34 (1984) 939–944.

[72] J.R. Cockrell, M. Folstein, Mini mental state examination (MMSE), Psychopharmacology 24 (1988) 689–692.

[73] R.M. Crum, J.C. Anthony, S.S. Bassett, M. Folstein, Population-based norms for the mini-mental state examination by age and educational level, Journal of American Medical Association 269 (1993) 2386–2391.

[74] M. Unser, The Gallery at wavelet.org. <http://www.wavelet.org/phpBB2/gallery.php> (accessed 21.10.06).