Proceedings of the 4th International
IEEE EMBS Conference on Neural Engineering
Antalya, Turkey, April 29 - May 2, 2009

Combining Multichannel ERP Data for Early Diagnosis of Alzheimer's Disease

Metin Ahiskali, Robi Polikar
Electrical and Computer Eng.
Rowan Univ., Glassboro, NJ USA

John Kounios, Deborah Green
Dept. of Psychology
Drexel Univ., Philadelphia, PA USA

Christopher M. Clark
Dept. of Neurology
Univ. of Pennsylvania,
Philadelphia, PA USA

Abstract – **As the average age of our population increases, the prevalence of Alzheimer's Disease (AD), the most common form of dementia, has grown sharply. Current diagnosis of AD primarily uses longitudinal clinical evaluations and/or invasive lumbar punctures for CSF analysis, available only at specialized hospitals, which are generally outside of financial and geographical reach of most patients. We expand on our previous work and describe an ensemble of classifiers based approach that combines decision and data fusion techniques for the early diagnosis of AD using event related potentials (ERP) obtained in response to different audio stimuli. In this contribution, we specifically examine various feature set combinations, obtained from different EEG electrode locations and in response to different stimulus tones to illustrate the accuracy of such a system for AD diagnosis at the earliest stage on a clinically significant cohort size of 71 patients.**

I. Introduction

Alzheimer's disease (AD) represents one of the greatest health risks to our aging population. A neurodegenerative disorder, AD is caused by neuronal death due two misfolded proteins, β-amyloid and hyperphosphorylated–τ, which cause plaques and neurofibrillary tangles, respectively. Symptoms of AD include a gradual loss of memory, motor skills, and cognitive impairment. While the disease affects an average of 2% of those under 65, the prevalence doubles every five years [1;2]. Because of the debilitating effects of the disease on the patient, the emotional stress on the family or caregivers, and steep financial toll on society, AD has become a major health concern.

While there is currently no treatment that can stop the progression of AD, recent pharmacological developments, such as acetylcholinesterase inhibitors or glutamate blockers can slow the development of AD [3]. However, such drugs require that the disease be diagnosed at the earliest stage possible, which is a significant challenge. Currently the definitive diagnosis for AD require analyzing brain tissue under the microscope for the presence of plagues and tangles, a method only available during an autopsy. A pre-mortem diagnosis is commonly done though repeated longitudinal clinical evaluations, which include multiple memory tests of the subject, and interviews of both subject and their caretakers, and/or the highly invasive lumbar puncture for the analysis of the cerebrospinal fluid for the presence of β-amyloid and hyperphosphorylated–τ. These diagnoses can provide 90% accuracy; however, they are only available at specialized clinics of major health centers, and are extremely expensive. At local clinics and hospitals, where most patients seek care due to geographic or financial restrictions, the accuracy of diagnosis is estimated as 75%, even with frequent patient monitoring [4] .

Hence, a reliable, non-invasive, and cost-effective approach is needed which can be made available to local clinics. Electroencephalogram (EEG), the only brain monitoring technology that can provide reasonable time resolution, may help provide such an approach. Previous studies have shown that the event related potentials (ERP), which are time-locked averages of the EEG recorded in response to certain stimuli using the so-called oddball paradigm can provide diagnostically useful information for AD. Specifically, a decrease in amplitude and increase in latency of the P_{300} component of the ERP, a positive peak that occurs 300 ms after the stimulus, has been linked to cognitive decline and AD [5-9]. Various signal processing approaches on the raw EEG or the P_{300} has been conducted since then, which verified the presence of a statistical correlation, albeit a weak one that has mixed success in patient specific diagnosis [10-13]. Previous studies have shown that discrete wavelet coefficients of the ERPs, and not that of just the P_{300} components, are more beneficial in patient specific AD diagnosis, particularly when the ERPs in response to different types of stimuli are combined [14-17]. In this contribution, we show that a decision fusion based approach to data fusion, obtained through stacked generalization, that combines location specific information obtained from different electrodes with stimulus-type specific information provides an even further improvement in diagnostic accuracy. We have also increased our cohort to 71 subjects for a more clinically significant cohort size.

II. Experimental Setup

A. The Oddball Paradigm and ERP Acquisition

The ERPs are acquired using an auditory oddball paradigm protocol. 19 electrodes were used, embedded in an elastic cap with two reference electrodes on the ears according to the International 10-20 electrode placement system (Figure 1). Each electrode was kept below an impedance of 20kΩ to ensure proper connection with the scalp. Each subject was tested for 30 minutes with approximately three minutes of rest for every five minutes of testing. The actual data recording was preceded with a one minute practice session. Binaural audiometric thresholds were first determined using a 1 kHz tone presented to both of the subject's ears at 60 dB above the subjects hearing threshold to prevent any bias based on the hearing of the patients.

This work is supported by Neuronetrix, Inc., NIH Grant No. P30 AG10124 - R01 AG022272, PA Dept. of Health Grant SAP4100027296, and by National Science Foundation Grant No ECS-0239090.

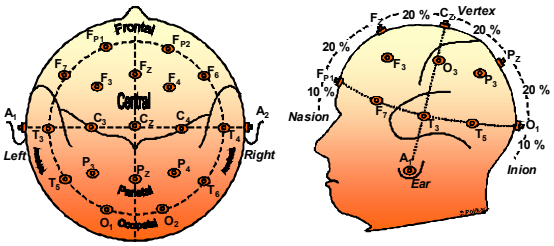

Figure 1. 10-20 International system of electrode placement

1000 random stimuli were presented to each subject with 65% non-target stimuli of 1 kHz tones, 20% of target stimuli of 2 kHz tones, and 15% of novel sounds. A random inter-stimulus interval of 1.0~1.3 seconds was used. Normal and target stimuli were delivered in 100 ms long burst with 5ms onset and offset envelopes. The novel stimuli were environmental sounds, 200ms long, each presented once to ensure novelty. The subjects were instructed to press a button each time the target tone was presented, and do nothing for non-target or novel tones. Time-locked averages of stimulus specific responses were computed from raw EEG to obtain the ERPs, with each record starting 200ms before the stimulus and ending 800ms after the stimulus. Recordings containing artifacts were rejected by an EEG technician. Remaining ERPs were amplified, digitized at 256 Hz, and notch filtered at 59-61Hz. The pre-stimulus baseline was subtracted from the entire ERP, resulting in 1 s duration of 256 samples per stimulus type, per channel, per patient.

B. Patient Cohort

The primary goal in recruiting the AD cohort was to include subjects who were in the earliest stages of the disease. Along with a clinical assessment of each subject, the cognitive level of each subject was measured using the Mini Mental State Exam, a standardized test that accesses orientation, attention, immediate and short-term recall, language, and ability to follow written and verbal commands. The test is scored on a scale of 0 to 30, 30 being cognitively normal, and 0 being vegetative state. A 19 or lower score is considered cognitive impairment. The average MMSE score for the normal (control) cohort was 29, whereas that of AD cohort was 25, indicating that the AD cohort was in their earliest stages of the disease. Inclusion criteria for AD group was satisfying the NINCDS-ADRDA criteria [18] 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 was evidence of any other central nervous system damage, or use of sedatives, anxiolytic or antidepressants within 48 hours of ERP acquisition. Final cohort included 71 subjects, 34 with AD (average age 74) and 37 cognitively normal (average age 76).

III. METHODS

A. Feature Extraction

Event Related Potentials are non-stationary signals whose frequency content change over time. In order to extract time-localized frequency band specific information from such signals, a time-frequency representation, such as the discrete wavelet transform (DWT) is appropriate. The DWT decomposes the signal into frequency sub-band using a series of successive highpass and lowpass filters and subsampling at each level. The outputs of the highpass filters are the detail (DWT) coefficients, whereas the outputs of lowpass filters are the approximation coefficients. Using an eight coefficient long Daubechies-4 wavelet, the ERPs were decomposed into following frequency bands:

$\boldsymbol{d}_1$:64~128 Hz (N=132) $\boldsymbol{d}_5$: 4 ~ 8 Hz (N=14)
$\boldsymbol{d}_2$: 32 ~ 64 Hz (N=69) $\boldsymbol{d}_6$: 2 ~ 4 Hz (N=10)
$\boldsymbol{d}_3$: 16 ~ 32 Hz (N=38) $\boldsymbol{d}_7$: 1 ~ 2 Hz (N=8)
$\boldsymbol{d}_4$: 8 ~ 16 Hz (N=22) $\boldsymbol{a}_7$: 0 ~ 1 Hz (N=8).

Since the primary information in the ERPs are known to be in the sub 8 Hz band, we focused on the decomposition levels 5-7, resulting in four frequency bands d_5, d_6, d_7, and a_7, to explore as various feature sets. We also look at the responses to both the target tones and novel sounds, obtained from each electrode. We then constructed a data fusion system for combining information from different electrode locations, stimulus types and frequency bands using a ensemble-based decision fusion approach.

B. Ensemble Based Systems

We implement an ensemble of classifiers based decision fusion approach for automated classification. An ensemble system consists of a group of classifiers trained on different subsets of training data or different feature spaces to generate different decision boundaries. Classifiers then make different errors on different instances, and a strategic combination of these classifiers can aid in reducing the total error [19]. Most ensemble approaches fall into one of two categories: classifier selection or classifier (decision) fusion [20;21]. In *classifier selection*, each classifier is trained to become an expert in some local area of the feature space. Given some instance **x**, the classifiers trained with data closest to the vicinity of **x** will make the final decision. In *classifier* (*decision*) *fusion* all classifiers are trained over the entire feature space. A classifier combination rule then merges these individual classifiers to form an "expert" to aid in an overall system performance increase. Note that *classifier fusion* traditionally means the combination of classifiers for classification improvement, where all classifiers are trained on the same data source. This is different than *data fusion*, where data from different sources are combined. In the approach described below, we use stacked generalization as a decision fusion approach, which creates an ensemble of classifiers for each data source. An ensemble based expert is therefore created for each data source (for example, for each electrode, stimulus type and frequency band combination). The decisions of these experts are then further combined using an ensemble combination rule to achieve decision fusion based data fusion.

C. Decision Fusion Using Stacked Generalization

The underlying concept in stacked generalization is to use a meta-classifier to confirm or correct what has already been learned by a group of preliminary (Tier-1) classifiers. Instances occupying a certain region of the feature space may be more likely to be misclassified by certain classifiers than others. Such a trend can be learned by mapping the outputs of an ensemble of classifiers to their true labels.

In Wolpert's stacked generalization, an ensemble of classifiers are first created, whose outputs are used as inputs to a second level meta-classifier to learn the mapping between

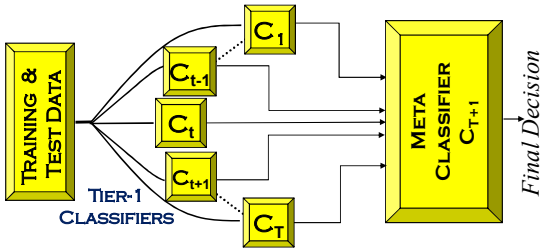

Figure 2. Block diagram of the standard stacked generalization.

between the ensemble outputs and the correct class labels [22]. The general block diagram of stacked generalization can be seen in Figure 2.

D. Augmented Stacked Generalization

We use a modified version of stacked generalization, called the augmented stacked generalization (ASG) by augmenting the Tier-1 classifier outputs with the *original data* used to train them, before training the meta-classifier. Such a process enriches the intermediate feature space used by the meta-classifier, thus aiding in overall system performance [23]. A two stage training process is implemented, where initial training begins with a K-fold cross validation on the training dataset (in the original feature space). This output is then utilized as the input (training data in the intermediate space) for the meta-classifier.

ASG starts with dividing the training data of length $N = 71$ into K (roughly) equal blocks, each of length N/K. K was chosen as 5 in our implementation. Each classifier in the ensemble, C_1 through C_T, is trained K times, using K-1 blocks of the training data (~56 subjects). For each such training, one block of data is not seen by Tier-1 classifiers. The classifier outputs for each block not seen during training (~15 patients) are then augmented with the original data used to train that classifier, which creates the training data for the meta-classifier (C_{T+1}). The original training labels are used in the training process, allowing the meta-classifier to determine – and correct – poorly performing classifiers.

In the intermediate training stage of ASG, the original Tier-1 classifiers are discarded, and all N instances of the training data are collected. The Tier-1 classifiers are then retrained on the entire training data subset. During the testing stage, a given test instance is sent to the Tier-1 classifiers. The output from these classifiers (augmented with the original feature vector) is then sent as an input to the MetaClassifier. The output of the MetaClassifier then constitutes the final decision and output of the system. A pseudocode of this process is shown in Figure 3.

E. Data Fusion

The output from the ASG algorithm is a decision-fusion-based expert for each data source. Training an ensemble of such experts (each with data from different sources) creates a decision fusion based data fusion approach.

There are many methods available for combining ensemble of classifiers' outputs. In this study, we evaluated the sum and majority voting schemes due to their simplicity and reported general superiority [24;25]. In-depth discussion of the mathematics behind different combination rules, as well as their advantages and disadvantages can be found in [24;26].

Augmented Stacked Generalization (ASG)
Inputs

- Training data $\boldsymbol{S} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, with correct labels $y_i \in \Omega,\ \Omega = \{\omega_1, \cdots, \omega_C\}$;
- Supervised classifiers for **BaseClassifier** and a **MetaClassifier**, which can be of the same type.
- Number of classifiers T to be generated

Initial Training: Train Tier-1 classifiers:
Divide $\boldsymbol{S}$ into K blocks of size N/K, i.e. $\boldsymbol{S}_1$, $\boldsymbol{S}_2$,...,$\boldsymbol{S}_k$,
Do for k= 1,2,...,K

- Train classifiers C_1,...,C_T on K-*1* blocks of data;
- Test classifiers C_1,...,C_T on the K^{th} block;
- Obtain continuous-valued output of each classifier for each class in [0 1] range, indicating each classifiers support for each class

End Do loop
Train meta-classifier C_{T+1}:

- Form the augmented feature-space by concatenating Tier-1 classifier outputs with the original $\mathbf{x}_n$ used to train Tier-1 classifier;
- Train MetaClassifier C_{T+1} using N instances of augmented data along with their corresponding class labels ω_1,...,ω_C;

Intermediate training

- Retrain Tier-1 classifiers C_1,...,C_T on the entire $\boldsymbol{S}$

Testing: Given an unlabeled instance **x'**

- Obtain and concatenate the outputs of C_1,...,C_T,
- Augment the outputs of C_1,...,C_T with the original **x'** to obtain the intermediate space feature vector
- Obtain the output of C_{T+1} as the predicted label for **x'**.

Figure 3. Pseudocode of the ASG algorithm

IV. RESULTS

The diagnostic accuracies for top performing 16 individual feature sets (experts), as obtained by the stacked generalization, are shown in Table 1, all of which were in the 0 – 4 Hz range. All performance figures are averages of 100 independent trials obtained by random shuffling of the data. 95% confidence intervals are also provided for all averages. The multi-layer perceptron (MLP) was used as the base classifier (as it provides continuous estimates of support for each class, useful for ASG and necessary for the sum rule), with a single hidden layer of 10 nodes and 0.01 error goal, chosen with cross validation based prior experience on this data. The naming convention in Table 1 is as follows: [tone] [electrode] [frequency band], e.g., NC_Z12 represents ERP obtained in response to Novel tone, at CZ electrode, decomposed to 1-2 Hz band (i.e., using DWT d_7 coefficients).

The decisions obtained using these feature sets are then combined using the sum and majority vote rules, whose results are shown in Table 2. Three feature set combinations were created for the final ensemble decision based on prior knowledge, electrode location and spectral diversity.

FS_1:NC_Z12+NC_Z24+NP_Z24+NT_812+NP_Z24+$TF_{P2}12$+ TP_Z01+TF_812+TP_324
FS_2:NP_Z12+NP_Z24+NC_Z24+NT_812+$TF_{P2}12$+NC_Z12
FS_3:$TF_{P2}12$+TF_812+TP_Z01+NC_Z12+NP_Z24+NT_812+NP_Z12

TABLE 1 - INDIVIDUAL FEATURE SET PERFORMANCES

Feature Set	Average (%)	Best Trial (%)
NP_Z12	70.11±1.19	88.67
NP_Z01	69.41±1.85	77.13
NC_Z12	69.13±1.36	78.00
TP_Z12	67.55±1.65	81.02
TP_Z24	66.87±0.86	89.89
TF_812	66.51±1.87	82.48
NF_Z24	66.23±1.73	75.14
NC_Z24	64.70±1.23	88.19
NT_812	63.99±1.32	86.44
$TF_{P2}12$	63.97±1.59	79.11
NP_Z24	62.97±1.16	73.69
TP_324	62.58±1.77	79.66
TP_Z01	62.54±1.01	81.33
TP_312	62.54±1.95	79.66
TC_Z24	62.43±1.25	79.55
NO_Z12	61.79±1.73	79.40

TABLE 2. - ENSEMBLE SYSTEM PERFORMANCE

SUM		
	Avg (%)	Best(%)
FS_1	84.51±1.8	94.70
FS_2	81.30±1.7	82.99
FS_3	82.44±2.0	87.03

MAJORITY VOTING		
	Avg (%)	Best(%)
FS_1	74.33±1.3	82.14
FS_2	71.78±2.1	86.60
FS_3	73.15±2.2	82.00

Table 2 indicates that ASG based ensemble performance exceeds that of both community clinic diagnostic accuracy, as well as individual feature set performances. The sum rule on the first feature set combination performed the best overall, attaining an average performance of 84.51 ±1.8% (average of 100 trials), with the single best performance (of the 100 trials) being 94.70%. These are significantly improved (3~5%) performance figures compared to our previous efforts, where we have looked at standard boosting based approaches as reported in [14-17].

V. CONCLUSIONS

The primary goal of this study was to investigate the effectiveness of decision fusion based data fusion approach in combining ERP data from different electrode locations and stimulus types used in EEG data collection for the early diagnosis of Alzheimer's disease. These preliminary results indicate that there is indeed complementary information in different data sources, which can be extracted and synergistically combined using the described approach. The diagnostic accuracy of this approach, reaching approximately 84%, significantly exceeds that of community clinics, and approaches even that of expert neurologists using the NINCDS-ADRDA criteria. We should note however that the gold standard used in this work was the data labeled by expert neurologists, and hence the performance figures simply reflect the ability of the approach to match the diagnosis of these experts. The actual performance of the approach to diagnose the disease can be slightly below or above the reported numbers.

REFERENCES

[1] Alzheimer's Society Statistics Available online at: http://www.alzheimers.org.uk/site/scripts/documents_info.php?categoryID=200120&documentID=341, Last accessed :1-19-2009

[2] C. P. Ferri, et al., "Global prevalence of dementia: a Delphi consensus study," *The Lancet*, vol. 366, no. 9503, pp. 2112-2117, 2005.

[3] Alzheimer's Society Drug treatments for Alzheimer's disease Available online at: http://www.alzheimers.org.uk/factsheet/407, Last accessed :1-19-2009

[4] A. Lim, et al., "Clinico-neuropathological correlation of Alzheimer's disease in a community-based case series," *Journal of the American Geriatrics Society*, vol. 47, no. 5, pp. 564-569, 1999.

[5] J. Polich, C. Ladish, and F. E. Bloom, "P300 assessment of early Alzheimer's disease," *Electroencephalogr. Clin. Neurophysiol.*, vol. 77, no. 3, pp. 179-189, May1990.

[6] S. Yamaguchi, et al., "Event-related brain potentials in response to novel sounds in dementia," *Clinical Neurophysiology*, vol. 111, no. 2, pp. 195-203, 2000.

[7] 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.

[8] 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," *Journal of Clinical Neurophysiology*, vol. 24, no. 5, pp. 405-412, 2007.

[9] J. Polich, "Updating P300: An integrative theory of P3a and P3b," *Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128-2148, 2007.

[10] 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.

[11] A. A. Petrosian, D. V. Prokhorov, W. Lajara-Nanson, and R. B. Schiffer, "Recurrent neural network-based approach for early recognition of Alzheimer's disease in EEG," *Clin. Neurophysiol.*, vol. 112, no. 8, pp. 1378-1387, Aug.2001.

[12] M. Karrasch, et al., "Brain oscillatory responses to an auditory-verbal working memory task in mild cognitive impairment and Alzheimer's disease," *International Journal of Psychophysiology*, vol. 59, no. 2, pp. 168-178, Feb.2006.

[13] T. Demiralp, A. Ademoglu, Y. Istefanopulos, 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] R. Polikar, "Bootstrap - Inspired Techniques in Computational Intelligence," *IEEE Signal Processing Mag.*, vol. 24, no. 4, pp. 59-72, 2007.

[20] L. I. Kuncheva, "Switching between selection and fusion in combining classifiers: An experiment," *IEEE Tran. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 32, no. 2, pp. 146-156, 2002.

[21] K. Woods, W. P. J. Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Trans. Pattern Analysis and Machine Intelli.*, vol. 19, no. 4, pp. 405-410, 1997.

[22] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241-259, 1992.

[23] S. Shin, K. Lee, and S. Kilic, "Ensemble Prediction of Commercial Bank Failure Through Diversification of Input Features," *AI 2006: Advances in Artificial Intelligence*, vol. 4304, 2006, pp. 887-896.

[24] J. Kittler, M. Hatef, R. P. W. Duin, and J. Mates, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, 1998.

[25] G. Fumera and F. Roli, "Linear Combiners for Classifier Fusion: Some Theoretical and Experimental Results," *4th Int. Workshop on Multiple Classifier Systems*, Lecture Notes in Computer Science, vol. 2709, 2003, pp. 74-83.

[26] G. Fumera and F. Roli, "Linear Combiners for Classifier Fusion: Some Theoretical and Experimental Results," *4th Int. Workshop on Multiple Classifier Systems*, Lecture Notes in Computer Science ,T. Windeatt and F. Roli, Eds. vol. 2709, 2003, pp. 74-83.