Evaluation of a Neural-Network Classifier for PET Scans of Normal and Alzheimer's Disease Subjects

J. Shane Kippenhan, Warren W. Barker, Shlomo Pascal, Joachim Nagel, and Ranjan Duara

Wien Center for Memory Disorders, Mt. Sinai Medical Center, Miami Beach, Florida and Departments of Biomedical Engineering, Radiology, Neurology, University of Miami, Coral Gables, Florida

The value of PET as an objective diagnostic tool for dementia may depend on the degree to which abnormal metabolic patterns can be detected by quantitative classification methods. In these studies, a neural-network classifier based on coarse region of interest analyses was used to classify normal and abnormal FDG-PET scans. The performance of neural networks and of an expert reader were evaluated by cross-validation testing. When the "abnormal" class was represented by subjects with clinical diagnoses of "Probable Alzheimer's," the areas under the relative-operating-characteristic (ROC) curves were 0.85 and 0.89 for the neural network and the expert reader, respectively. When testing with abnormal subjects represented by "Possible AD" cases, ROC areas for both the network and the expert were 0.81. The neural network out-performed discriminant analysis. It is concluded that PET has potential for the detection of abnormal brain function in dementing diseases, and that the combination of neural networks and PET is a useful diagnostic tool. Despite the low-resolution "view" afforded the neural network, its performance was nearly equivalent to that of an expert reader.

J Nucl Med 1992; 33:1459–1467

Positron emission tomography (PET) scan studies in dementia have shown so-called "typical" patterns of abnormality, such as bilateral parieto-temporal hypometabolism, asymmetrical hypometabolism and predominantly frontal hypometabolism (*1–9*) as shown in Figure 1. These "typical" patterns appears in many (usually advanced) cases of different neurological disorders, and it is often possible to show significant differences in PET scan patterns on a group basis (e.g., differences in mean values of particular regions, or of ratios of mean values to a reference region, for a given disorder), but reliable case-by-case classification of subjects remains difficult. Disorders such as Parkinson's dementia and normal pressure hydrocephalus can produce metabolic patterns that are supposedly "typical" of Alzheimer's disease (AD) (*10,11*).

Because of the large numbers of regions of interest (ROIs) in a typical PET data base, identification of either discrete abnormalities or patterns of abnormality for a patient group compared to a control group has posed some statistical challenge. PET studies are expensive and involve radiation exposure, and because PET itself has not been a commonly available methodology, relatively few subjects in patient and control groups have been studied. Furthermore, intrinsic variability and methodological artifacts result in considerable inter- and intrasubject variability. Multicolinearity or interdependence between brain regions also presents problems in investigating regional differences between groups. These problems have been addressed to some extent by normalization of regional values to some reference region(s) (*9,12*), by the use of multivariate approaches (*13*), by discriminant-function analysis (*14*) and by a "scaled sub-profile" model (*15*). These different methods of analysis have value for identifying single or groups of brain regions that have the greatest differences between subject groups (best discriminators) or for identifying a distinctive profile of regional function that characterize a disorder. AD and other memory disorders are heterogeneous, however, and there is growing evidence of multiple sub-types within the AD disease category (*16–19*). This evidence indicates the need, which has not yet been met, for a PET analysis method which has the ability to recognize and employ multiple discriminating profiles that will serve to identify certain disorders on PET scans.

PET scan classification methods can vary from interpretation by a human reader to more automated methods, as shown in Figure 2. Certain artifacts, such as those produced by a lateral tilt of the brain, can be suspected and taken into account more readily by a human expert than by currently-available computerized methods. Needless to say, the performance of a human reader will depend on that reader's level of expertise and experience in reading PET scans.

To evaluate a quantitative classification method adequately, *cross-validation* studies must be performed. This involves "training" the classifier with one group of subjects,

Received Oct. 29, 1991; revision accepted Mar. 25, 1992.

For reprints contact: J. Shane Kippenhan, PhD, Wien Center for Memory Disorders, Mount Sinai Medical Center, 4300 Alton Rd., Miami Beach, FL 33140.

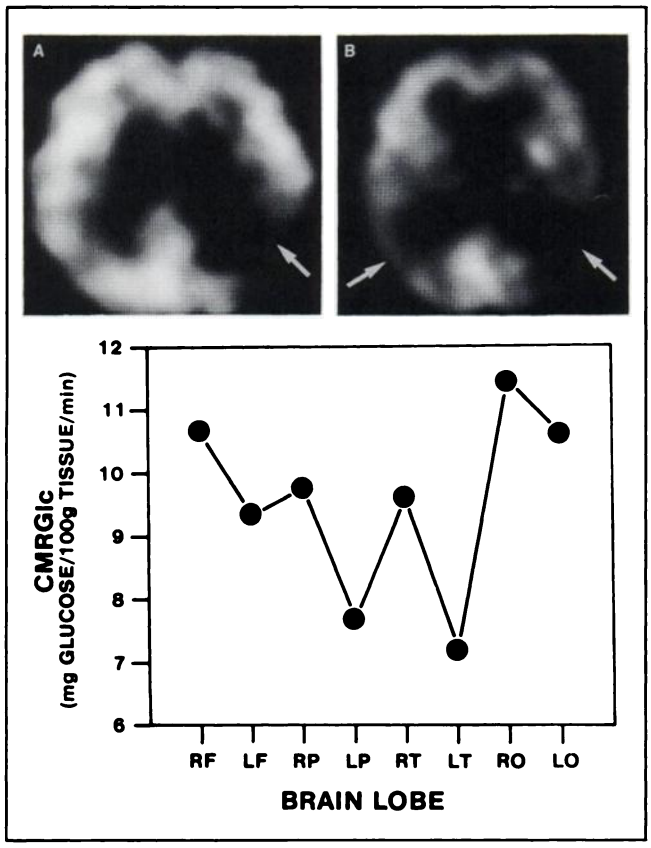

FIGURE 1. PET image illustrating parieto-temporal asymmetries and bilateral parieto-temporal hypometabolism. The plot gives results of a ROI analysis of the PET study shown above, i.e., CMRglc in the four bilateral lobes of the brain (right and left frontal, parietal, temporal, occipital).

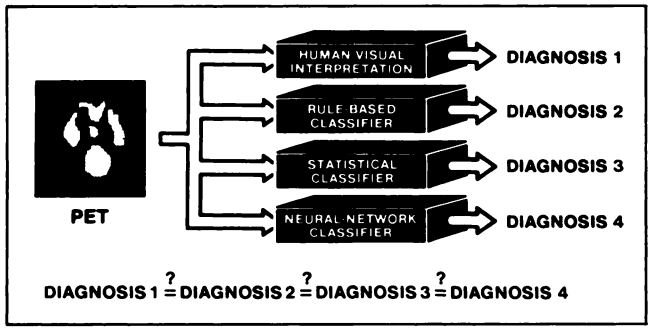

FIGURE 2. A selection of PET classification methods. Information from a PET study (either the images themselves or the results of quantitative analyses) can be used to "diagnose" abnormalities. Intuitive visual interpretations by trained experts can be powerful, but are inherently subjective. Certain "rules" regarding patterns of deficits or assymetry may be formalized to constitute a "rule-based classifier." Quantitative, statistical versions of these types of rules, or approaches such as discriminant analysis, may be used to form a "statistical classifier." Alternatively, neural networks may be trained to indicate abnormalities. The relevant question is: given the same PET scan, will the different methods agree?

and testing it on a separate group, which will measure a classification method's ability to perform on new and independent data sets. *Re-substitution* experiments, in which testing is performed on the group used to "train" the classifier, can be used to demonstrate theoretical limits of the classifier's performance as the number of experimental subjects increases (*20*), but they do not give realistic estimates of a classifier's performance in a practical setting.

The introduction of a group of computational algorithms known collectively as "artificial networks" has stimulated great interest within the field of pattern recognition (*21–23*). In these algorithms, individual processing elements, analogous to biological neurons, receive weighted averages of inputs from other processing elements. As in biological networks, a transfer function is applied to this weighted average, and the results are fed to other processing elements. The recent development of methods by which connection weights can be adjusted so that networks can "learn," by example, how to classify patterns, has made this technique particularly valuable. The introduction of the generalized delta rule (*21*) for use in semi-linear networks enabled the realization of computer programs implementing a multilayer neural network (the back-propagation network) that could perform impressive feats of "learning."

Though neural network training is strictly a "supervised learning" process, the learning is essentially by example, with no guidance from the user as to the criteria to employ. The network is allowed to learn what it "believes" to be the most important discriminating features, and to weigh those features appropriately for best classification performance. The weight vectors associated with individual hidden units can quite legitimately be thought of as "feature detectors," since the weighted-average input to each hidden unit represents a covariance or correlation-type calculation. In the sense that it applies multivariate profiles to PET data, the back-propagation network approach is conceptually similar to the methods of others (*14,15,19*). The nonparametric and nonlinear aspects of neural networks, however, offer potential advantages.

Neural networks are beginning to find applications in many fields, including the field of medical imaging (*24–29*). A neural-network classification system for fluorodeoxy-glucose (FDG) PET scan data is described here, and its applicability in separating normal from abnormal FDG-PET scans is evaluated.

MATERIALS AND METHODS

Patients with dementia or memory disorders were recruited for brain imaging studies at the Wien Center for Alzheimer's Disease and Memory Disorders, Mount Sinai Medical Center, Miami Beach. Normal young and elderly subjects were also recruited from the local community. Recruiting procedures are described in detail elsewhere (*8*).

Resting-state (supine, awake, eyes closed and blindfolded in a quiet, darkened room) PET scans were obtained using a PETT V scanner (*30*) (seven simultaneous slices, 15 mm apart, with

inplane and axial image resolution of 15 mm FWHM). Patients were injected with 3–5 mCi of [^{18}F]FDG, and scans were obtained 30 min later for a length of time sufficient to obtain 2×10^6 counts in the highest count slice. "Arterialized" venous blood was collected in order to measure plasma radioactivity and glucose (*31*). Regional cerebral metabolic rate of glucose (rCMRglc) values were calculated using standard rate constants, a lumped constant of 0.42 and an operational equation (*31*). Data were analyzed for 67 ROIs in the brain, using previously-published methods (*13,32–34*). For each region, the average metabolism in absolute values of rCMRglc in mg/100 g/min was determined. Values were also calculated for 12 larger bilateral lobular regions and 4 bilateral lobar regions (frontal, parietal, temporal and occipital).

Classification performances were evaluated for two groups of subjects. All subjects had been clinically diagnosed according to current NINCDS-ADRDA criteria (*35*), and the clinical diagnosis was used as a reference. Each group contained two classes: "AD" subjects and age-equivalent normal subjects. The AD class was represented, in the first group, by subjects with a clinical diagnosis of "Probable AD", and in the second, by subjects with "Possible AD" as defined by NINCDS-ADRDA criteria (it was expected that PET-based diagnosis would be somewhat more difficult for the Possible AD group). Approximately half of the patients diagnosed as Possible AD had met all the criteria for Probable AD, except for being insufficiently cognitively impaired to be labeled as demented. The remaining Possible AD patients were those who were demented but had other medical conditions that could independently produce some mental impairment, so as to make the diagnosis of AD less certain. Table 1 summarizes the composition of the two groups.

Three classification methods were compared. The first method consisted of classification by a human expert. An expert reader (RD), who was blind to individual clinical diagnoses, examined each PET scan for signs of abnormality and assigned a grade of abnormality from 0 through 5 (0 = completely normal, 1 = questionable deficit present, 2 = mild deficits, 3 = moderate deficits, 4 = severe deficits, 5 = severe widespread deficits). A threshold-type decision criterion was then applied to each subject. The results of the human expert were compared to the results of two quantitative classifier methods, both of which were trained with eight-dimensional patterns (eight lobar metabolic values) resulting from ROI analyses of individual subjects. One quantitative method was a discriminant analysis technique (*36*), as implemented in the SAS statistical package (*37*), in which the discriminant function obtained for a "training set" was applied to patterns within a "testing set". The SAS procedure employed an optimization strategy which used either linear or quadratic discriminant analysis, depending on the results of tests of the intra-class and pooled covariance matrices (*38*). The second quantitative method was the back-propagation artificial neural network. Both cross-validation and re-substitution studies were performed for the quantitative methods.

Comparisons of the three methods were made on the basis of "relative-operating-characteristics" (ROC) analyses (*39–42*), in which the area under the ROC curve was used as the figure of merit. The ROC area measures a diagnostic system's performance at several different settings of the decision criteria, and is a more complete representation of a diagnostic system's performance than, for example, the report of a single pair of sensitivity and specificity values. It can be shown (*42*) that the area under the curve corresponds to the probability of a correct response in a two-alternative forced choice test, in which a classifier is presented with one sample of each of the two possible alternatives (in this case, normal or abnormal), and is forced to say which is which.

Figure 3 is a conceptual representation of the neural-network classification system. ROI data, based on rCMRglc in the eight (four right and four left) lobes served as an eight-dimensional input to the neural network. Neural network training was performed using back-propagation techniques described elsewhere (*21,22*). By presenting examples of each class (in this case, results of ROI analyses of normal and AD PET scans) at the input layer, comparing the calculated output of each output-layer unit with the target values for that class, and then adjusting the internal weights so that the calculated outputs would then be closer to the target values, the network learned to define appropriate decision boundaries within the input space. Once a network was trained in this way, "unknown" patterns were classified by presenting input patterns at the input layer. For the two-class problem, a single output unit indicated the classification, according to a selected threshold criterion. The network was trained so that the output of this unit was "high" (close to 1) for normal patterns and "low" (close to 0) for abnormal patterns. Target values used for training were thus either 1 or 0, for normal and abnormal subjects, respectively. For one iteration, the entire group of examples was presented, and the error at the output layer was

TABLE 1
Composition of the Two Groups Used to Test Classification Performance

	Group 1	Group 2
"AD" class:	"Probable AD"	"Possible" AD
N	41	39
Age	70.9 ± 8.8 (range: 53–93)	73.6 ± 9.4 (range: 51–96)
Mini-Mental status exam score	15.0 ± 7.3	19.0 ± 8.0
Gender (M, F)	21, 20	18, 21
"Normal" class:	Age-Equivalent Normals	Age-Equivalent Normals
N	50	50
Age	67.7 ± 8.9 (range: 50–84)	67.7 ± 8.9 (range: 50–84)
Gender (M, F)	25, 25	25, 25

Note: "Normals" include some subjects with small MR lesions.

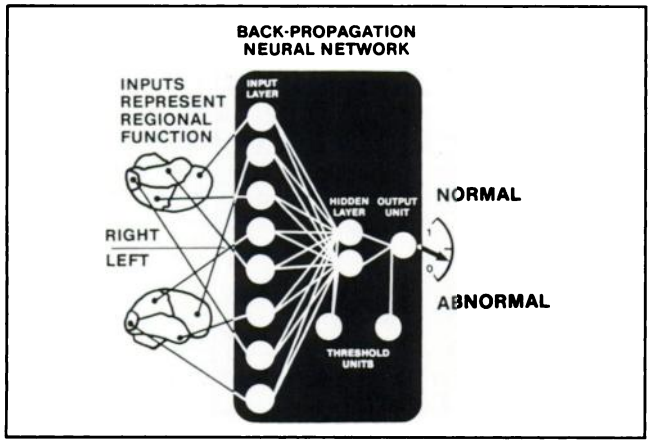

FIGURE 3. A conceptual representation of the PET classification system based on a neural-network classifier. Values representing metabolism in various regions of the brain serve as input patterns. Examples of patterns from previously-classified normal and abnormal subjects can be used to perform "supervised" training of the neural network. The classification performance of the network can then be tested by presenting metabolic patterns of new subjects. The number of units in the "hidden layer" can be varied to optimize the cross-validation performance.

"propagated" back through the network according to the generalized delta learning law (*21*).

The major architectural parameters of a back-propagation neural network are: the number of hidden layers, the number of hidden units within each layer and the number of training iterations. These parameters all affect the capacity of the network to "generalize" when performing classification. Although some theoretical guidelines for optimizing these parameters with respect to a particular application can be found in the literature (*22,43–45*), this type of optimization is still somewhat of an open question. It is advisable to use the simplest architecture possible and to train for no longer than necessary, since overtraining can cause a network to "memorize" its training set and degrade its performance on the testing set. The network may learn classification "rules" which apply specifically to the training patterns and are not generally applicable. Neural networks with a single hidden layer were optimized with respect to the number of hidden units and to training duration. Optimization procedures were performed as described elsewhere (25). Briefly, classification performances, as judged by ROC areas in cross-validation testing, were evaluated for different combinations of number of hidden units and number of training iterations. Overtraining was considered to occur when average cross-validation ROC areas began to decrease with increasing training. The number of hidden units at which ROC areas no longer increased with an increasing number of hidden units was considered to be a number sufficient for the data under consideration.

Normal controls were randomly divided into two groups of equal size, and each group was then paired with an abnormal group, thus forming two independent data sets which could serve as training-testing pairs. In order to balance the number of normal and abnormal subjects in *training sets* (so as to eliminate any learning bias), a number of patterns from the smaller class were represented more than once. In order to make the cross-validation results as general as possible (i.e., to reduce the results' dependence on any special properties of a given training-testing combination), two different cross-validation configurations were used to obtain each averaged ROC curve. The normal controls used for training in the first configuration were used for testing in the second configuration, and vice versa. The cross-validation results from these two configurations were averaged to obtain final ROC curves. For neural-network experiments, training was repeated several times for each configuration, each time with random network initialization in order to eliminate any potential bias attributable to particular initial conditions. The order in which subjects within a training set were presented for training was also randomized.

ROC curves were constructed by determining pairs of true-positive-ratio and false-positive-ratio values at various settings of decision criteria for each method. For the expert, this was accomplished by selecting different thresholds of assigned abnormality grades (from 0 to 5) for the classification criteria. For the neural network, ROC curves were similarly computed by selecting different thresholds for the output units of networks which were trained to indicate abnormality on a scale of 0 to 1. For discriminant analysis, points on the ROC curve were collected by choosing a range of prior probabilities (from 0 to 1) for the discriminant procedure.

Because of the nature of the sigmoid transfer function of the network's processing units, it is usually necessary to pre-process the input data. The data should be scaled by an arbitrary constant chosen so that the input values are "small" (absolute values less than about 2). Another pre-processing option is to de-mean each input pattern, i.e., to subtract the mean value of each n-dimensional pattern from each of the n components. The inherent assumption here is that the mean value (indicating overall level of metabolism) is not as important to the classification process as are the relative differences among the individual regional function values. While it is not always advantageous to remove information from input patterns, it *can* be beneficial if the information removed is misleading or has little value. Experiments were performed for two pre-processing methods in order to quantitatively compare these methods. For one group of experiments, the mean was removed from each pattern, as described above, to form *zero-mean* patterns. For the second group of experiments, *non-zero-mean* patterns were formed by simply scaling the metabolic values so that their range was between 0 and 1.

RESULTS

The different methods were used to classify subjects in Group 1 ("Probable AD" versus age-equivalent normals), with the results shown in Figure 4. Since the specificity is the complement of the false-positive ratio represented on the abscissa in Figure 4, one can determine the sensitivity and specificity for various strengths of criteria (more "strict" or more "lenient") directly from the ROC curve. At a specificity of 80% (0.2 false-positive ratio), for instance, one can see that the sensitivity of both the expert reader and the neural network was in the range of 80%–85%, while the sensitivity of discriminant analysis was in the range of 65%–70%. The same methods were applied to subjects in Group 2 ("Possible AD" versus age-equivalent normals). As shown in Figure 5, the ROC curves for the neural network and the expert nearly overlap one

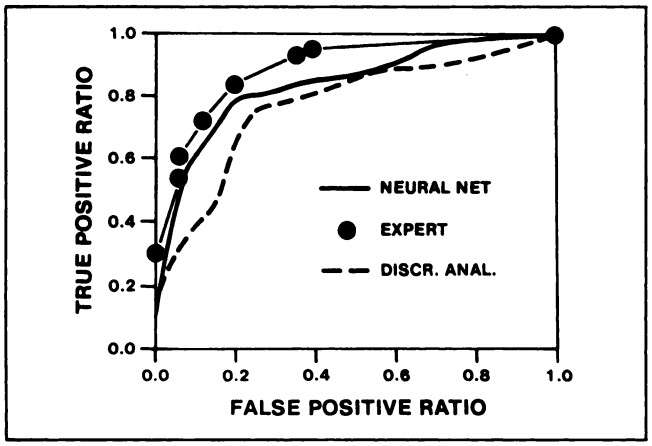

FIGURE 4. ROC curves illustrating classification performance within Group 1. Different points on the neural-network curve were determined by continuously varying the output unit's decision threshold. The neural-network curve shown is an average of ROC curves from forty trials (20 trials for each of two cross-validation configurations). Different points on the expert "curve" were determined by selecting different decision thresholds on the expert's 0–5 abnormality scale. In order of decreasing true-positive ratios, the points above correspond to thresholds of 0.0, 0.5, 1.0, 2.0, 3.0, 3.5, 4.0 and 5.0. Points on the discriminant analysis curve were determined by choosing a range of prior-probability values. The discriminant analysis curve was the average of results for two cross-validation configurations.

another. The discriminant analysis curve shows a lower sensitivity for nearly all values of false-positive-ratio.

Early experiments indicated that training-set size was an important factor in the generalizing capabilities of both quantitative methods. The influence of variations over the available range of training-set size appeared to outweigh any effects attributable to composition (i.e., whether the training set contained Probable AD or Possible AD subjects). In order to use the largest possible number of training samples while maintaining the independence of testing sets, training sets for classifiers to be tested on Probable AD sets included Possible AD subjects, and vice versa. In fact, experiments indicated that there was very little difference (variation in average ROC area was 0.01) between the case in which the abnormal groups for training sets consisted of equal ratios of Possible AD and Probable AD subjects and the case in which abnormal training groups was composed entirely of either Possible AD or Probable AD subjects. The cross-validation results presented here are for the case in which networks were trained on Group 1 and tested on Group 2, and vice versa. For re-substitution tests, networks trained with either Group 1 or Group 2 were also tested on that same group.

The ROC areas for both subject groups are summarized in Table 2. The neural-network ROC values shown in Table 2 are mean values. Neural-network cross-validation results were based on forty different training/testing experiments (twenty experiments for each of two cross-validation configurations), resulting in standard deviations of 0.012 for Group 1 and 0.018 for Group 2. For the cross-validation results shown in Table 2, networks with four hidden units (8-4-1 networks) were trained for just 40 iterations. These were optimal training parameters, as determined by the optimization procedure described earlier.

The results of the mean-removal comparison experiments are shown in Table 3. Results are shown for both quantitative methods for each of two types of data representation: zero-mean and non-zero-mean (as described earlier). Within Group 1, removing the mean resulted in slightly higher ROC areas for the neural network, and slightly lower ROC areas for discriminant analysis. For Group 2, removing the mean made little or no difference for either discriminant analysis or neural networks. Neural-network training times for non-zero-mean experiments were longer (400–500 iterations), than those for zero-mean experiments.

Selected weight vectors associated with the hidden units of networks that were trained to distinguish normal from

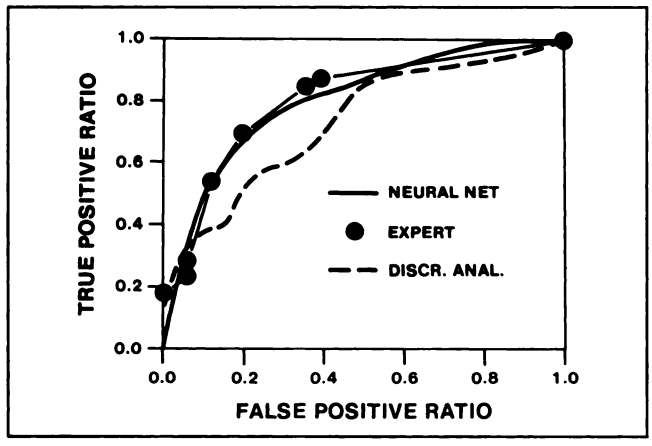

FIGURE 5. ROC curves illustrating classification performance within Group 2. Points on the curves were computed in the same manner as were points on the curves in Figure 4.

TABLE 2
Classification Performance of Various Classification Methods*

Method	Group 1 (Probable AD vs. Age-Equiv. Normal)	Group 2 (Possible AD vs. Age-Equiv. Normal)
Expert Reader	0.89	0.81
Neural Network (cross-validation)	0.85	0.81
Neural Network (re-substitution)	0.98	0.97
Discr. Analysis (cross-validation)	0.80	0.74
Discr. Analysis (re-substitution)	0.92	0.92

* Each value represents the area under the ROC curve for a given classification method.

TABLE 3
Classification Performance for Two Different Data Preparation Methods*

Method	Group 1 (Probable AD vs. Age-Equiv. Normal)	Group 2 (Possible AD vs. Age-Equiv. Normal)
Neural Network (zero-mean)	0.85	0.81
Neural Network (non-zero-mean)	0.82	0.80
Discr. Analysis (zero-mean)	0.78	0.74
Discr. Analysis (non-zero-mean)	0.80	0.74

* Each value represents the area under the ROC curve.

Note: Training times for non-zero-mean experiments were longer (400–500 iterations) than those for zero-mean experiments (40 iterations).

Probable AD PET scans are presented in Figure 6. These vectors represent the most distinctive and heavily-weighted "abnormal-detecting" patterns from groups of trained networks. The weight vectors presented here result from training with non-zero-mean patterns, which corresponds, of course, to the customary method by which human experts observe PET images, i.e., without mean removal.

DISCUSSION

The results of this work suggest that PET has a notable capacity for discriminating between normal and AD subjects, and that the back-propagation neural network is a useful classification tool. As indicated by ROC-based performance evaluations within test groups of different diagnostic difficulty, the neural network's performance was better than that of discriminant analysis and comparable to that of an expert PET reader's performance, despite the low-resolution image data (one value per lobe) provided to the network. The nonlinear and nonparametric nature of the neural network apparently allowed it to be a more robust classification procedure than those based on traditional statistical methods.

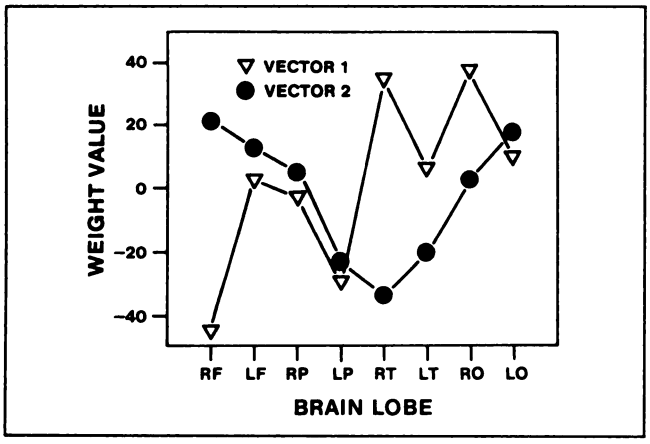

FIGURE 6. The two most important "discriminating profiles" used by the neural nets. Shown above are weight vectors of hidden units of networks trained on Group 1. The neural network has incorporated and combined some "typical Alzheimer's disease" features, particularly asymmetry and left-parietal hypometabolism, into its feature detectors. Note that the combination of the two weight patterns allowed a trained network to detect frontal asymmetry in *either* direction: right-side-*higher*-than-left or right-side-*lower*-than-left.

As expected, classification accuracy was higher within Group 1 than within Group 2 (Table 2). In general, patterns within the diagnostically more difficult group (Group 2) would be less descriptive of the "AD" group, thereby decreasing classification performance. The results of the re-substitution experiments show, however, that it is possible to *separate* normal from abnormal subjects with almost complete accuracy in both groups. This implies that higher accuracy for both groups could be obtained with larger training sets.

Comparison between the results obtained using zero-mean data versus those obtained with non-zero-mean data served to demonstrate, at least for this subject group, the relative *unimportance* of the mean value in discriminating between normal and abnormal PET scans. In general, removing the mean from the data presented to the quantitative classifiers either made no difference in classification performance, or made it *easier* to distinguish between the two groups. The single exception was a slight degradation in discriminant-analysis performance for Group 1, as shown in Table 3. For the neural network experiments, results with zero-mean values were either equal to or slightly higher than results with non-zero-mean values. Also, shorter training times indicated that it was easier to separate the two classes when using zero-mean values.

The results of classification within Group 1 are lower than those reported by Friedland et al. for a similar group of subjects studied with FDG-PET (*46*). There are at least two possible explanations for this difference: differences in PET-camera resolution, and differences in the method for collecting input-function blood samples. The metabolic values obtained with the Scanditronix camera used by Friedland et al. [with a 6 mm FWHM (*47*)] can differ significantly from those obtained from the same subject with a low-resolution camera (*48*). In addition, Friedland et al. performed arterial blood collection, rather than arterialized venous collection. The latter has been cited in more recent literature as a potentially significant source of error (*49*).

It should be remembered that there are three major issues to be considered in a PET-based classification system: (1) the intrinsic diagnostic power of PET imaging; (2) the quality of the image analysis; and (3) the classification method. Each of these matters will influence the performance of a classification system. Although the focus here is primarily on classification methods, it should be remembered that poor performance in either of the other two areas will compromise the classification results.

It can be seen from the weight patterns (feature vectors) shown in Figure 6 that the networks incorporated and combined several patterns of asymmetry and hypometabolism into their feature detection process. These feature vectors can be thought of as representing the relative effect of individual input variables while others are held fixed. These vectors, however, should be interpreted carefully. They represent profiles that have been adjusted to serve as discriminating profiles for abnormal PET scans on a *group* basis. Metabolic patterns, representing individual subjects, which were presented to the neural network and "matched up" with one or more of the prominent aspects of *one or both of* these weight vectors, were "judged" abnormal according to the degree of matching. An increase in the number and/or extent of matching profile characteristics corresponded to an increased indication of abnormality.

Several aspects of these profiles deserve mention. Vector 1 served as a mechanism for detecting patterns of prominent left-impaired asymmetry in parietal, temporal and occipital regions. These patterns, combined with sparing of the occipital and temporal regions (relative to other regions) and pronounced hypometabolism in the left-parietal lobe, were strong indicators of abnormality. Vector 2 served to detect occipital and temporal asymmetry in the *opposite direction*, i.e., right-side hypometabolism (though not as pronounced as in vector 1), combined with sparing of the occipital and frontal regions relative to other regions, and hypometabolism in the parietal and temporal regions. Frontal asymmetry was also an abnormal indicator, particularly in vector 1, which shows right-side-*lower*-than-left asymmetry. Vector 2 shows frontal asymmetry in the opposite direction, which allowed trained networks to detect frontal asymmetry in *either* direction. It can be seen that the combination of these two weight patterns allowed trained neural networks to detect quite a rich variety of abnormal indicators.

The ROC curve represents the performance at several different settings of the particular decision criteria. The area under the curve is the "only performance measure available that is uninfluenced by decision biases and prior probabilities, and it places the performances of diverse systems on a common, easily interpreted scale" (*40*). The area values presented above can be compared with values from the literature (*40*), which describe the diagnostic performances of various medical imaging techniques, such as the detection of brain lesions on CT (A = 0.97), on radionuclide scanning (A = 0.87) and the detection of adrenal disease (0.93 for CT, 0.81 for ultrasound).

The ability of this study to fully evaluate the diagnostic capability of neural networks was limited by several considerations. Although, for a PET study, this group of subjects was fairly large, it was small enough to impose limitations in two senses. First of all, the neural network's "past experience," in each evaluation, was represented only by the subjects in the training set. The expert reader's training, of course, was not limited to the data sets used to train the quantitative classifiers, but was based on knowledge gained from a professional lifetime's worth of experience. In addition, the low-resolution "view" that the neural network had of each PET study certainly represented a significant handicap.

Another factor which may have influenced the results is the fact that, if the expert reader is considered as an expert system in the same sense as were the quantitative classifiers, part of the expert's training was performed on the "testing set." Some of the expert's conclusions regarding trends of asymmetry, etc. may actually be rather specific for this group. In terms of the ROC values presented here, this represents an additional "handicap" on the quantitative classifiers, as compared to the expert reader. When evaluating the *quantitative* classifiers, the training and testing sets were completely independent.

Another limitation in evaluating the classification methods described here stems from the assumption that the clinical diagnoses are accurate. Postmortem studies have shown confirmation of clinical diagnoses for AD cases to average about 80% (*50–52*). While all abnormal subjects in this study presumably have an organic brain disorder, they may have, in approximately 20% of the cases, a disease other than AD. Another factor that may degrade specificity is the heterogeneity in pathological findings in AD. This heterogeneity may eventually produce several different "metabolic types" of AD. Thus, the standard by which the methods' performances are measured is itself somewhat uncertain. These limitations would particularly apply to the results from Group 2, whose "AD" diagnoses were less certain than those in Group 1.

The possible existence of metabolic sub-types noted previously could serve to help explain the higher accuracy of neural networks as compared to discriminant analysis. As the weight vector analyses have shown, the neural-network approach enables the identification of *more than one* characteristic metabolic profile, which is appropriate when a single disease may be manifested by more than one metabolic pattern.

The combination of PET and neural networks appears to be an objective and useful diagnostic tool for AD, and would appear to be well-suited for structure-function-based classification in other diseases as well. It should be noted that the current study did not involve *differential* diagnosis. Future work would include training with examples of more than one disease category.

Artificial neural networks can be used to model anatomical/functional disorders, since their architecture and processing modes are similar to those of biological networks. In the study presented here, patterns of regional function have been associated with clinical diagnoses. In a similar way, regional functional patterns can be associated with patterns in neuropsychological evaluations, which could lead to the discovery of patterns associated with very specific neurological or cognitive syndromes.

An image-based classification system could be easily

and effectively used to compare imaging modalities as well as procedure-specific parameters (tracer concentrations, number of counts to collect, etc.). Additional data from magnetic resonance scans and even data from neurological and psychological evaluations could be included to form the basis of a comprehensive expert system. Such a system, trained with the knowledge of human specialists, could be available wherever there was a computer and could be available on a continuous basis.

ACKNOWLEDGMENTS

Portions of the research reported here were funded under an agreement with the Aging and Adult Services Program Office, Department of Health and Rehabilitative Services, State of Florida. Additional support was provided by fellowship awards to J. S. Kippenhan from the University of Miami Graduate School, The Education and Research Foundation of the Society of Nuclear Medicine, The International Society for Optical Engineering and The Society for Imaging Science and Technology. Other support was from general research funds from Mt. Sinai Medical Center, Miami Beach, FL. We are indebted to Mr. J. Chang and Drs. A. Apicella and F. Yoshii for valuable assistance in performing and analyzing the PET scans.

REFERENCES

1. Foster NL, Chase TN, Fedio P, Patronas NJ, Brooks RA, DiChiro G. Alzheimer's disease: focal cortical changes shown by positron emission tomography. *Neurology* 1982;33:961–965.
2. Foster NL, Hansen MS, Siegel GJ, Kuhl DE. Medial and lateral temporal glucose metabolism in aging and Alzheimer's disease studied by PET. *Neurology* 1988;38(suppl 1):133.
3. Friedland RP, Budinger TF, Ganz E, et al. Regional cerebral metabolic alterations in dementia of the Alzheimer type: positron emission tomography with [^{18}F]fluorodeoxy-glucose. *J Comput Assist Tomogr* 1983;7: 590–598.
4. Duara R, Grady C, Haxby J, et al. Positron emission tomography in Alzheimer's disease. *Neurology* 1986;36:879–887.
5. Foster NL, Gilman S, Berent S, Morin EM, Brown MB, Koeppe RA. Cerebral hypometabolism in progressive supranuclear palsy studied with positron emission tomography. *Ann Neurol* 1988;24:399–406.
6. McGeer PL, Kamo H, Harrop R, et al. Positron emission tomography in patients with clinically diagnosed Alzheimer's disease. *Can Med Assoc J* 1986;134:597–607.
7. Kamo H, McGeer PL, Harrop R, et al. Positron emission tomography and histopathology in Pick's disease. *Neurology* 1987;37:439.
8. Loewenstein DA, Barker WW, Chang J, et al. Predominant left hemisphere metabolic dysfunction in dementia. *Arch Neurology* 1989;46:146–152.
9. Haxby JV. Resting state regional cerebral metabolism in dementia of the Alzheimer type. In: Duara R, ed. *Positron emission tomography in dementia*. New York: Wiley-Liss; 1990:93–116.
10. Schapiro MB, Grady C. Reductions in parietal/temporal cerebral glucose metabolism are not specific for Alzheimer's disease. *Neurology* 1990; 40(suppl 1):152.
11. Friedland, RP. 'Normal'-pressure hydrocephalus and the saga of the treatable dementias. *JAMA* 1989;262:2577–2581.
12. Duara R, Grady C, Haxby JV, et al. Human brain glucose utilization and cognitive function in relation to age. *Ann Neurol* 1984;16:702–713.
13. Yoshii F, Barker WW, Chang JY, et al. Sensitivity of cerebral glucose metabolism to age, gender, brain volume, brain atrophy, and cerebrovascular risk factors. *J Cereb Blood Flow Metab* 1988;8:654–661.
14. Powers WJ, Perlmutter JS, Videen TO. Accuracy of PET for detecting Alzheimer's disease. *J Nucl Med* 1990;31:730.
15. Moeller JR, Strother SC, Sidtis JJ, Rottenberg DA. Scaled subprofile model: a statistical approach to the analysis of functional patterns in positron emission tomographic data. *J Cereb Blood Flow Metab* 1987;7:649–658.
16. Haxby JV, Duara R, Grady CL, Culter NR, Rapoport SI. Relations between neuropsychological and cerebral metabolic asymmetries in early Alzheimer's disease. *J Cereb Blood Flow Metab* 1985;5:193.
17. Martin A, Brouwers P, Lalonde F, et al. Towards a behavioral typology of Alzheimer's patients. *J Clin Exp Neuropsychol* 1986;8:594.
18. Grady CL, Haxby JV, Horwitz B, et al. A longitudinal study of the early neuropsychological and cerebral metabolic changes in dementia of the Alzheimer type. *J Clin Exp Neuropsychol* 1988;10:576.
19. Grady CL, Haxby JV, Schapiro MB, Kumar A, Friedland R, Rapoport SI. Heterogeneity in dementia of the Alzheimer type (DAT): subgroups identified from cerebral metabolic patterns using positron emission tomography (PET). *Neurology* 1989;39(suppl1):167–168.
20. Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans Pattern Anal Machine Intelligence* 1991;13:252–264.
21. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, the PDP Research Group. *Parallel distributed processing volume 1*. Cambridge: MIT Press; 1986:318–364.
22. Lippman RP. An introduction to computing with neural nets. *IEEE ASSP Magazine* April, 1987;4–22.
23. Pao Yoh-Han. *Adaptive pattern recognition and neural networks*. New York: Addison-Wesley; 1989.
24. Kippenhan JS, Nagel JH. Diagnosis and modelling of Alzheimer's disease through neural network analyses of PET studies. *Proc 12th Ann Int Conf IEEE/EMBS* 1990;12:1449–1450.
25. Kippenhan JS, Nagel JH. Optimization and evaluation of a neural-network classifier for PET scans of memory-disorder subjects. *Proc 13th Ann Int Conf IEEE/EMBS* 1991;13:1472–1473.
26. Boone JM, Gross GW, Greco-Hunt V. Neural Networks in radiologic diagnosis: I. introduction and illustration. *Invest Radiol* 1990;25: 1012–1016.
27. Gross GW, Boone JM, Greco-Hunt V, Greenberg B. Neural networks in radiologic diagnosis: II. interpretation of neonatal chest radiographs. *Invest Radiol* 1991;25:1017–1023.
28. Floyd CE, Bowsher JE, Munley MT, Tourassi GD, Baydush AH Coleman RE. Neural network for quantitative reconstruction of SPECT images [Abstract]. *J Nucl Med* 1991;32:936.
29. Sun H, Mazoyer BM. Simulations of a Hopfield neural network based on cerebral glucose metabolism regional correlations measured by positron tomography [Abstract]. *J Cereb Blood Flow Metab* 1991;11(suppl 2): S371.
30. Ter Pogossian M, Mullani N, Hood J. Design considerations for a positron emission transverse tomograph (PETT V) for imaging of the brain. *J Comput Assist Tomogr* 1978;2:149–154.
31. Phelps M, Huang S, Hoffmann E, Selin E, Sokoloff L, Kuhl D. Tomographic measurement of local cerebral glucose metabolic rate in humans with (F-18) 2-fluoro-2-deoxy-D-glucose: validation of method. *Ann Neurol* 1979;6:371–388.
32. Duara R, Margolin R, Robertson-Tchabo E, et al. Cerebral glucose utilization, as measured with positron emission tomography in 21 resting healthy men between the ages of 21 and 83 years. *Brain* 1983;106: 761–775.
33. Duara R, Grady C, Haxby JV, et al. Human brain glucose utilization and cognitive function in relation to age. *Ann Neurol* 1984;16:702–713.
34. Duara R, Gross-Glenn K, Barker W, et al. Behavioral activation and the variability of cerebral glucose metabolic measurements. *J Cereb Blood Flow Metab* 1987;7:266–271.
35. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. 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* 1984;34:939–944.
36. Rao CR. *Linear statistical inference and its applications*. New York: Wiley; 1973.
37. *SAS/STAT guide for personal computers*, version 6 edition. Carry, NC: SAS Institute Inc.; 1985:83–110.
38. Morrison DF. *Multivariate statistical methods*. New York: Mc-Graw Hill; 1976.
39. Swets JA, Pickett RM. *Evaluation of diagnostic systems*. New York: Academic Press; 1982.
40. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988; 240:1285–1293.
41. Metz CE. ROC methodology in radiologic imaging. *Invest Radiol* 1986; 21:720–733.
42. Green DM, Swets JA. *Signal detection theory and psychophysics*. New York: Wiley; 1966. Reprinted by Krieger, Huntington, NY, 1974.

43. Baum EB. On the capabilities of multilayer perceptrons. *J Complexity* 1988;4:193–215.
44. Mirchandani G. On hidden nodes for neural nets. *IEEE Trans Circuits and Systems* 1989;36:661–664.
45. Mehrotra KG, Mohan CK, Ranka S. Bounds on the number of samples needed for neural learning. *IEEE Trans Neural Networks* 1991;2: 548–558.
46. Friedland RP, Horwitz B, Grady C, et al. An ROC analysis of the diagnostic accuracy of PET with FDG and x-ray computed tomography in Alzheimer's disease [Abstract]. *J Cereb Blood Flow Metab* 1989;9(suppl 1):S566.
47. Daube-Witherspoon ME, Green MV, Holte S. Performance of Scanditronix PC1024-7B PET scanner [Abstract]. *J Nucl Med* 1987;28:607–608.
48. Grady CL. Quantitative comparison of measurement of cerebral glucose metabolic rate made with two positron cameras. *J Cereb Blood Flow Metab* 1991;11:A57–A63.
49. Carson RE. Precision and accuracy considerations of physiological quantitation in PET. *J Cereb Blood Flow Metab* 1991;11:A45–A50.
50. Kukull WA, Larson EB, Reifler BV, Lampe TH, Yerby MS, Hughes JP. The validity of 3 clinical diagnostic criteria for Alzheimer's disease. *Neurology* 1990;40:1364–1369.
51. Sulkava R, Haltia M, Paetan A, Wikstrom J, Palo J. Accuracy of clinical diagnosis in primary degenerative dementia: correlation with neuropathological findings. *J Neurol Neurosurg Psychiatry* 1983;46:9–13.
52. Mendez MF, Mastri A, Frey HF, Thomas A. Diagnostic trends in Alzheimer's disease: clinicopathological evidence in 383 cases from the Ramsey dementia brain bank [Abstract]. *Neurology* 1990;40(suppl 1):177.