# Bayesian Classification of FMRI Data: Evidence for Altered Neural Networks in Dementia

**John Burge** LAWNGUY@CS.UNM.EDU
Department of Computer Science, University of New Mexico, Albuquerque, New Mexico
**Vincent P. Clark** VCLARK@UNM.EDU
Departments of Psychology and Neuroscience, University of New Mexico, Albuquerque, New Mexico
The MIND Institute, Albuquerque, New Mexico
**Terran Lane** TERRAN@CS.UNM.EDU
**Hamilton Link** HAMLINK@COMCAST.NET
**Shibin Qiu** SQIU@UNM.EDU
Department of Computer Science, University of New Mexico, Albuquerque, New Mexico

## Abstract

The alterations in functional relationships among brain regions associated with senile dementia are not well understood. We present a machine learning technique using dynamic Bayesian networks (DBNs) that extracts causal relationships from functional magnetic resonance imaging (fMRI) data. Based on these relationships, we build neural-anatomical networks that are used to classify patient data as belonging to healthy or demented subjects. Visual-motor reaction time task data from healthy young, healthy elderly, and demented elderly patients (Buckner et al. 2000) was obtained through the fMRI Data Center. To reduce the extremely large volume of data acquired and the high level of noise inherent in fMRI data, we averaged data over neuroanatomical regions of interest. The DBNs were able to correctly discriminate young vs. elderly subjects with 80% accuracy, and demented vs. healthy elderly subjects with 73% accuracy. In addition, the DBNs identified causal neural networks present in 93% of the healthy elderly studied. The classification efficacy of the DBN was similar to two other widely used machine learning classification techniques: support vector machines (SVMs) and Gaussian naïve Bayesian networks (GNBNs), with the important advantage that the DBNs provides candidate neural anatomical networks associated with dementia. Networks found in demented but not healthy elderly patients included substantial involvement of the amygdala, which may be related to the anxiety and

agitation associated with dementia. DBNs may ultimately provide a biomarker for dementia in its early stages, and may be helpful for the diagnosis and treatment of other CNS disorders.

**Introduction**

Dementia is a progressive organic mental disorder characterized by impairments in short- and long-term memory combined with impairments in abstract thinking, judgment, disturbances of higher cortical function, personality change and increased anxiety and agitation (Volicer and Hurley 2003). Dementia's cause is most often Alzheimer's disease (AD), although other illnesses can produce dementia as well. Such global changes in behavior are likely to involve the disruption of activity across a variety of brain regions. Post-mortem histology has shown that AD involves medial temporal regions first and foremost, with a progressive loss of associated regions through the progression of the illness, leading finally to death (Braak, Del Tredici and Braak 2003). However, these histological methods cannot show the functional relationships among brain regions found to be affected or unaffected by neurodegeneration. Indeed, it is not entirely clear at what stage of degeneration the functionality of brain regions is fully impaired, nor to what degree functionality is compromised at intermediate stages of degeneration. Such information must be obtained from patients *in-vivo*, not only by observing the activity of individual brain regions during rest or in response to stimuli and behavioral responses, but also by observing how activity in one brain region affects others, and how this relationship changes as the behavioral symptoms of a disease manifest themselves. Information regarding such interactions among brain regions could prove useful for understanding the neural basis of dementia during intermediate stages of neurodegeneration, and may eventually lead to new methods for the diagnosis and treatment of this disorder.

Buckner et al. (2000) examined the neurophysiological and hemodynamic correlates of healthy aging and dementia by acquiring fMRI data during a visual-motor response task. The study consisted of three groups: 14 healthy young adults, 15 healthy adults and 11 demented adults. Subjects were presented

with flickering checkerboard stimuli and asked to press a button in response to stimuli. Using the generalized linear model (GLM) for analysis, they found qualitatively similar activation maps among all groups, but quantitatively differing amplitude responses in only a select few regions—such as the visual cortex. Over all, they found little quantitative differences among groups.

Using dynamic Bayesian networks (DBNs), a machine learning modeling technique that does not make the same linearity or independence assumptions as the GLM (or other widely used analysis techniques such as multivariate regression), we have developed a method for extracting causal[1] neuroanatomical networks differentiating activity among the three groups examined by Buckner et al. (2000). To validate the method, a classifier based on the DBNs was created and its over all accuracy was compared with two other popular machine-learning classifiers: support vector machines (SVMs) (Burges 1998) and Gaussian naïve Bayesian networks (GNBNs). (See the methods section for an overview on DBNs, GNBNs and SVMs.)

Both the DBNs and the GNBNs are Bayesian techniques that fall under the general category of Bayesian networks. However, the DBNs have the ability to explicitly model complex relationships among anatomical regions over time, whereas GNBNs collapse all temporal knowledge of a region into a few statistical values. While GNBNs are unable to illuminate meaningful temporal relationships, the simplicity of the model allows for efficient learning and GNBNs often make good classifiers.

In contrast, SVMs are not based on BNs, but are instead regression based-techniques that attempt to find optimal (possibly nonlinear) surfaces that separate sets of data points associated with different classes. For a general background on SVMs, we refer the reader to Burges (1998). Typically, SVMs map a problem into a higher dimensional domain in which standard quadratic programming techniques can be

[1] Relationships expressed in a DBN can be strictly causal (when explicitly constructed from causal processes) or correlational (when extracted from pure data). Traditionally, in Bayesian network research, we loosely refer to both types of relations as "causal". Following this nomenclature, we refer to "causal arcs" or "causal relations" extracted from data, though we acknowledge that, lacking additional experiments, the networks we demonstrate are only correlational.

used to find optimal linear separators (which result in nonlinear separators in the original domain). Like GNBNs, SVMs frequently make good classifiers, but they also do not readily provide useful information that could describe underlying neuroanatomical networks.

The enormous amount of information generated by an fMRI scan poses a significant obstacle to applying machine learning techniques. In order to reduce the size of the dataset while maintaining meaningful anatomical information, we averaged intensity data for voxels falling within each of the 150 anatomical ROIs as defined by the Talairach database (Lancaster et al. 2000)[2]. This results in a reduction from approximately 7 million spatial by temporal values per patient down to 70,000 values per patient—a reduction of two orders of magnitude.

Training the DBN on the processed data, we find that the classification accuracy of the DBN is competitive with both the SVM and GNBN and is capable of extracting meaningful neuroanatomical relationships describing qualitative and quantitative differences among groups. A method for measuring the confidence in the networks elucidated by the DBN is introduced and shows that, with high probability, many of the relationships reported in the neural anatomical networks are *not* due to random chance.

**Previous Work**

Much of the previous work in analyzing fMRI images uses statistical regression and hypothesis testing. A software tool commonly used to perform such analysis is SPM (Friston 2002). GLMs are frequently used to determine if a stimulus is correlated with voxel activity (Friston et al 1995). Derived from the GLM, the t-test has also been applied via SPM for fMRI analysis (Worsley & Friston 1995).

Generally, these techniques are employed on a voxel by voxel basis. However, Goutte et al. (1999) propose a method to cluster voxels with similar behavior into individual groups using both a K-means

[2] The database actually contains 151 anatomical regions, however, the precentral gyrus is duplicated, and one of the duplicates may be removed, resulting in 150 anatomical regions.

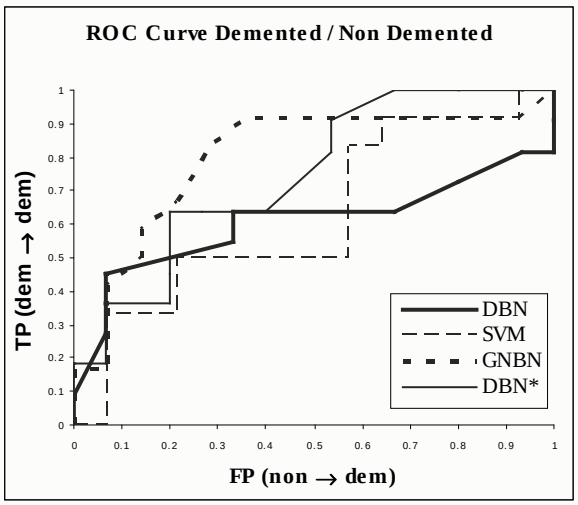



Figure 1. ROC curves for classifying demented vs. healthy patients. No single algorithm completely dominated although SVM was continually beaten by at least one other algorithm. '*' indicates a DBN with no parents (a multinomial model)

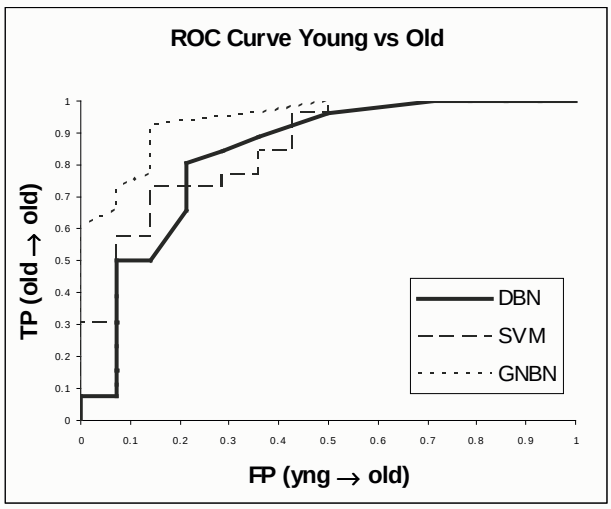



Figure 2. ROC curves for classifying age of patient. GNBN narrowly beat out both the DBN and SVM, which generally performed equally as well.

clustering algorithm and a hierarchical clustering algorithm that is capable of detecting clusters of voxels with correlations over varying time delays. The main purpose of their study was exploratory, as opposed to inferential, and they were able to successfully identify groups of similarly behaving voxels.

Discriminant analysis techniques (e.g., multivariate regression) are also commonly used to analyze neuroimaging data, and they share many modeling similarities with BNs. Like BNs, they test hypotheses on the predictive ability of independent random variables on dependant random variables. However, BNs tend to test predictive ability in an information theory context as opposed to a statistical analysis of variance approach typically employed by multivariate regression techniques. Using a multivariate regression on positron emission tomography (PET) data, Azari et al. examined the neurophysiological effects of dementia in elderly patients (Azari et al., 1993). They were able to correctly classify patients with mild or moderate dementia from healthy patients with 87% accuracy. While the subject population between Buckner et al.'s study and Azari et al.'s study are similar, Buckner et al.'s patients were only diagnosed with either very mild or mild demented and presents a more difficult classification that the mild or moderate dementia present in Azari et al.'s study.

Other Bayesian techniques have also been previously applied to fMRI data. Hojen-Sorensen et al. (2000) use Hidden Markov Models (HMMs) to learn a model of the activity within the visual cortex from visual stimuli. Using their technique, they were capable of reconstructing the visual stimulus presented to patients solely based on the patient's fMRI data.

The work most similar to ours is that of Mitchell et al. (2003). They used a Bayesian approach to analyze fMRI data with the goal of classifying instantaneous cognitive states of a patient such as 'reading a book' or 'looking at a picture'. They use three separate algorithms for analyzing the fMRI data: a GNBN classifier, SVMs and a K-Nearest Neighbor (KNN) classifier. They found that that KNN was uniformly outperformed by the SVM and GNBN, however, neither the SVM nor the GNBN clearly dominated the other.

Friston, Harrison and Penny (2003) have also employed Bayesian network modeling techniques to attempt reconstruction of activation networks. Like us, they employ dynamic (a.k.a. temporal) Bayesian networks, though they restrict their state transition function to bilinear functions of the current state and stimulus. We instead assume a multinomial transition model and do not constrain the prior distribution of hemodynamic response. Our more general model leads to a vastly more complex system identification problem that requires a number of approximations (see Methods section).

Most significantly, our work differs from previous work in that we are extracting meaningful networks differentiating groups of patients by explicitly modeling causal temporal relationships between neuroanatomical regions. Our primary focus is not that of classification, but learning how the relation between neuroanatomical regions changes between healthy and demented patients.

**Experimental Results and Discussion**

In this section, we describe the results of our experiments with the assumption the reader has some exposure to the Bayesian network terminology introduced in the Methods section.

| Experiment | Health | CI | Age | CI |
|---|---|---|---|---|
| SVM - Raw - Linear | 0.63 | 0.18 | 0.67 | 0.14 |
| SVM - Raw - Gaussian | 0.62 | 0.18 | 0.65 | 0.14 |
| SVM - Talairach - Linear | 0.65 | 0.18 | 0.78 | 0.12 |
| SVM - Talairach - Gaussian | 0.62 | 0.18 | 0.63 | 0.14 |
| DBN - Highlow2 - 5N - 4P | **0.73** | 0.17 | 0.65 | 0.14 |
| DBN - Highlow - 30N - 5P | 0.50 | 0.19 | **0.80** | 0.12 |
| GNBN - All Features | 0.61 | 0.18 | 0.62 | 0.15 |
| GNBN - Best Features | **0.73** | 0.17 | **0.80** | 0.12 |

Table 1. LOOCV accuracy and confidence intervals. 'Raw' and 'Talairach' indicate the SVM used individual voxel values or Talairach anatomical regions as input, respectively. 'Highlow2' and 'Highlow' indicate which dataset the DBN used. 5N and 30N indicate the setting for *numBestToKeep* and 4P and 5P indicate the setting for *numParents*. The highest accuracy results are bolded. The GNBN and DBN classifiers were able to achieve equivalent classification accuracies. See Methods section for parameter descriptions.

| numParents | Healthy | CI | Age | CI |
|---|---|---|---|---|
| 0 | 0.65 | 0.18 | 0.57 | 0.15 |
| 1 | 0.65 | 0.18 | 0.70 | 0.14 |
| 2 | 0.69 | 0.17 | 0.65 | 0.14 |
| 3 | 0.69 | 0.17 | 0.65 | 0.14 |
| 4 | **0.73** | 0.17 | 0.72 | 0.13 |
| 5 | 0.69 | 0.17 | **0.80** | 0.12 |
| 6 | 0.69 | 0.17 | 0.77 | 0.13 |
| 7 | 0.61 | 0.18 | 0.67 | 0.14 |

Table 2. LOOCV accuracy and confidence intervals for the DBN classifier with varying *numParents* parameter. The 'Healthy' column lists accuracy for classifying dementia and the age column lists accuracy for classifying age. The Healthy classifier used the *Highlow2* dataset with *numBestToKeep* equal to 5. The Age classifier used the *Highlow* dataset with *numBestToKeep* equal to 30. The best accuracy results are bolded.

Table 1 gives the classification accuracies for each of the three techniques we applied: DBNs, SVMs and GNBNs. In classifying healthy vs. demented patients, both the DBN and GNBN performed equally well, showing significantly higher classification accuracies than the SVM. However, the ROC curves given in Figures 1 and 2 clearly indicate that the classification behavior between the DBN and the GNBN was significantly different.

The DBN correctly classified 73% of elderly subjects as healthy or demented and of the healthy patients, the DBN correctly classified 93% of them. Thus, in discriminating healthy vs. demented patients, the DBN did well, but did better identifying healthy subjects than demented patients on a case-by-case basis. The GNBN also correctly classified 73% of elderly patients, however, only achieved an 84% accuracy rate on the healthy elderly patients. While the GNBN and DBN had identical overall classification accuracies, the DBN more accurately classified healthy elderly patients and the GNBN more accurately classified demented patients.

In overall accuracy for classifying the age of patients, all three models performed approximately equivalently. Again, the SVM and DBN were capable of classifying one of the classes, the young class, with higher accuracy at the cost of lower accuracy for the other class, the elderly. The GNBN showed

less bias for one class and achieved a more balance of classification accuracies between the two classes. This can be seen in the GNBN's clear dominance in the ROC curves given in Figure 2.

While the overall accuracy results between the GNBN and the DBN are largely similar, the information they are modeling on the patients is drastically different. The GNBN compresses all of the temporal behavior of the fMRI data into single valued means and variances whereas the DBN is explicitly modeling behavior over time. Initially, the classification ability of the DBN increases as more parents are added to each of the nodes representing neuroanatomical regions until model over fitting occurs (training accuracy reaches 100% after two parents are added). Table 2 details the improvement of leave one out cross validation (LOOCV) accuracies as the number of parents in each node are increased. This clearly indicates valuable temporal information exists in the fMRI time series that differentiates demented from healthy elderly individuals (as well as old from young).

It is worth noting the importance of the quantization method (a.k.a. amplitude decimation) used in the datasets. The three different quantization methods, *Highlow*, *Highlow2* and *Highlow4* each resulted in significantly different classification efficacies. The best classification of age was achieved when using the *Highlow* dataset which results in poor results when applied to classification of dementia. Likewise, use of the *Highlow2* dataset resulted in the highest accuracy for dementia, but when applied to age classification did poorly. The *Highlow4* dataset, while containing more information than the other datasets, unilaterally resulted in lower classification accuracies.

**Comparison of DBN Structure Between Groups**

We have compared our structural results for the demented vs. healthy DBN structure search with previous literature regarding putative functions of these regions. Table 3 lists the most predictive structures within each of the DBN classifiers as well as measures of confidence for those structures (see methods section for confidence analysis). For each of the predictive structures, the parents of the structures are also listed. If a structure is a parent of another structure, it indicates that the behavior of

| Five Best Predictors of Class | | | | | |
|---|---|---|---|---|---|
| Healthy Elderly | | | Demented Elderly | | |
| Structure | Parents | Confidence | Structure | Parents | Confidence |
| Gray Matter | Cuneus | 30.17 | BA 40 | Parietal Lobe | 9.17 |
| | Cingulate Gyrus | 39.38 | | Amygdala | 11.93 |
| | Rectal Gyrus | 24.58 | | Uvula of Vermis | 10.50 |
| | BA 34 | 9.18 | | Lateral Dorsal Nucleus | 7.64 |
| Left Cerebrum | Cuneus | 31.01 | Inf. Parietal Lobule | Inferior Parietal Lobule | 8.75 |
| | Cingulate Gyrus | 34.35 | | Uvula of Vermis | 11.19 |
| | BA 13 | 21.33 | | Amygdala | 11.33 |
| | BA 28 | 8.37 | | Inf. Occipital Gyrus | 8.52 |
| BA 4 | BA 4 | 16.99 | BA 7 | Inferior Parietal Lobule | 14.10 |
| | Occipital Lobe | 22.12 | | Declive | 17.00 |
| | Pyramis | 20.53 | | Amygdala | 12.75 |
| | Rectal Gyrus | 11.94 | | Inferior Semi Lunar Lobule | 8.24 |
| BA 6 | Gray Matter | 11.89 | Right Cerebrum | BA 31 | 12.61 |
| | Paracentral Lobule | 13.67 | | Insula | 14.65 |
| | Cerebellar Tonsil | 11.02 | | Amygdala | 10.55 |
| | Dentate | 6.12 | | Inferior Semi Lunar Lobule | 5.94 |
| Right Cerebrum | Right Cerebrum | 25.05 | Parietal Lobe | Parietal Lobe | 9.08 |
| | Cuneus | 31.70 | | Right Cerebellum | 10.26 |
| | Dentate | 19.82 | | Amygdala | 7.75 |
| | Uncus | 7.54 | | Lat. Dorsal Nucleus | 4.86 |

Table 3. The five most predictive structure networks found by the DBNs for classification of healthy elderly (left) and demented elderly (right) listed from highest to lowest BDE scores. The parents of each DBN network are given in the second column. The confidence for each network is the distance in the number of standard deviations of the network's BDE score from a mean BDE score calculated from a distribution of networks learnt from randomly permuted data. Notice that the structures with the highest BDE scores do not necessarily have the highest confidence measures. Confidence values are related to subject number, so are larger for healthy elderly.

the parent acts as a good predictor for the behavior of the other structure[3]. Together, the child and parent regions compose a *family*. These *families* form substructures within the DBN that indicate which neuroanatomical networks differentiate healthy patients from demented patients.

The finding that the only common good predictive structure between demented and non demented patients is the average over the entire right cerebrum, and includes the left cerebrum and all of gray matter for healthy subjects, suggests that dementia is associated with global dysfunction over multiple brain regions, rather than being associated with focal dysfunction in discrete brain regions.

[3] The parent/child relationship in Bayesian networks is roughly analogous to the independent/dependent relationship between variables in multivariate analysis.

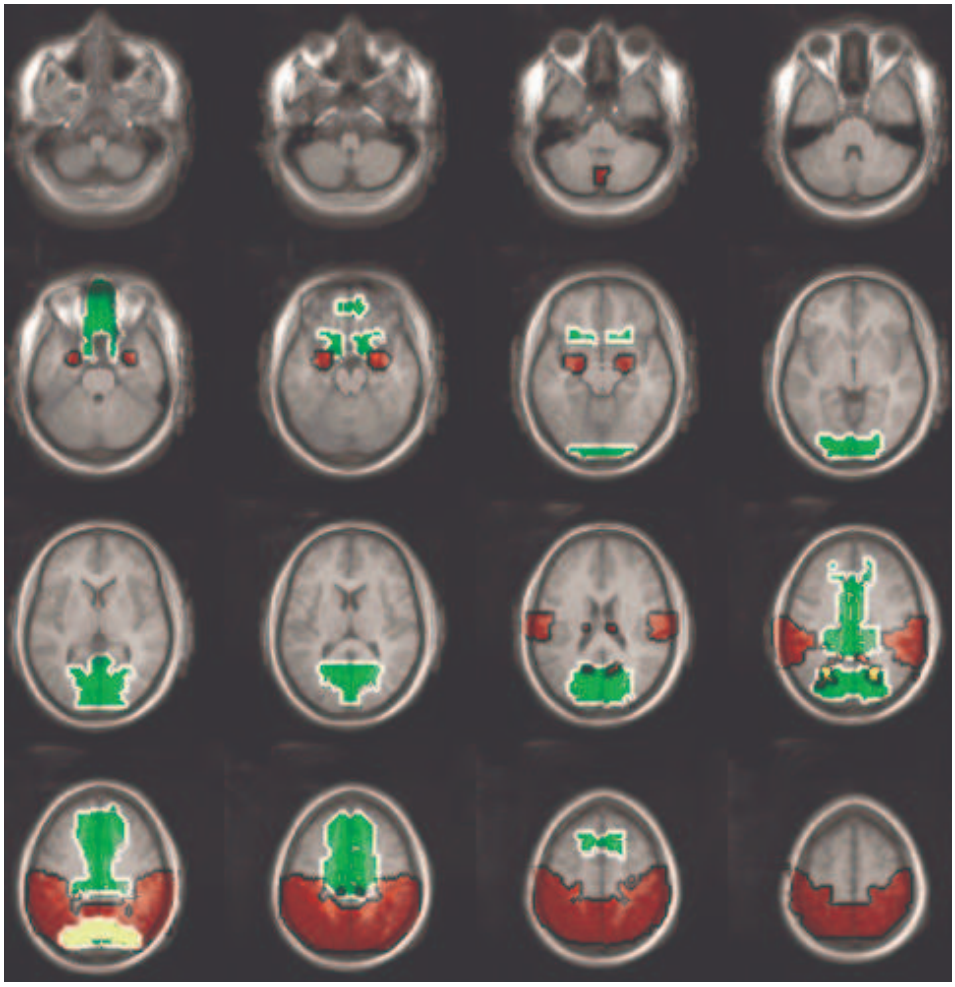


Figure 3. Visualization of the parent sets for the most predictive region in the demented patients (red, with black outline) and healthy elderly (green, with white outline), plotted onto a multi-subject averaged normalized images for illustration. Slices positioned from 56 mm below the origin to 64 mm above the origin. Regions overlapping demented and healthy subjects are shown in yellow. The most predictive structure for the healthy elderly was gray matter as a whole. The highlighted points indicate the union of the parents reported in the DBN search for this structure: the cuneus, cingulate gyrus, rectal gyrus and Brodmann area 34. The most predictive structure for the demented patients was Brodmann area 40. The highlighted points indicate the union of the parents for this node: the parietal lobe, amygdala, uvula of the cerebellar vermis and lateral dorsal nucleus of the thalamus.

The networks identified in most healthy elderly and some demented patients involved many visual, motor and attention regions already found to be involved in the visual-motor task by the GLM analysis of Buckner et al. (2000). The whole of gray matter, as well as left and right cerebrum separately, all have the cuneus as a parent, indicating that the behavior of the cuneus acts as a good predictor of the overall behavior of much of the brain during performance of this visual-motor task. The cuneus is a region of visual cortex located on the medial aspect of the occipital lobe superior to the calcarine fissure

and medial to the parieto-occipital fissure. This region was likely involved in identifying the visual stimuli presented in the study. Motor areas BA4 and BA6 were both found to be children in the healthy subjects. They are areas involved in initiating and coordinating responses to stimuli, and have parents that include other motor areas, as well as cerebellar regions likely involved in coordinating the motor response to stimuli, and perhaps other cognitive functions as well. The cingulate gyrus is a parent for both gray matter and the left cerebrum. Again, this is consistent with the role of the cingulate gyrus in response selection, attention and coordination of activity in other brain regions.

Other regions were found in healthy elderly that are not directly involved in the visual-motor task used, but that can be affected in dementia, and therefore may provide a better classification of healthy vs. demented elderly. These include the rectal gyrus, the uncus, BA 34, BA 28 and BA 13. The rectal gyrus was found to be a parent for both global gray matter and BA4 children. This region connects to the olfactory bulbs and is involved in olfaction, but may also be involved in emotional responses to stimuli (Clark et al. 2000, 2001; Clark 2002). The uncus, BA 34 and BA 28 regions are also involved in olfaction and emotion as a part of the limbic system, as well as memory encoding. Finally, BA 13 is a paralimbic region located within the insula. While the visual-motor task would not be expected to involve olfaction per se, the other cognitive functions of the limbic system could be involved, including emotion, motivation, memory and integrative processing. These functions are all disturbed in dementia, and one or more are likely to be evoked in some way by any stimulus or action. The finding that the networks formed by these regions are more likely to be found in healthy than demented subjects is reasonable given that progressive dementia leads to a loss of temporal and limbic structures initially.

The finding of a predominance of the amygdala as a parent for all 5 children associated with dementia, but for none of the healthy group, suggests that the amygdala exerts more control over the function of other brain regions in dementia relative to healthy ageing. This is interesting given that one predominant symptom of advanced dementia is agitation, which is a state of anxiety accompanied by

motor restlessness. Anxiety in dementia is often associated with feelings of impending danger, powerlessness, apprehension and tension. Physiological responses associated with agitation and anxiety include increased heart rate, altered respiration rate, sweating, trembling, weakness and fatigue[4]. This pattern of anxious behavior might be expected to involve the amygdala. Enhanced contribution of amygdalar activity to global brain function points to the possibility that these behavioral effects are related to specific alterations in brain function identified using the DBN.

One previous study has found an increased contribution of the amygdala to the function of other brain regions in dementia. Grady, Furey, Pietrini, Horwitz and Rapoport (2001) examined functional interactions between prefrontal and medial temporal brain areas during face memory using PET. They used a delayed face matching paradigm with delays from 1 to 16 s. Memory performance was correlated with amygdalar activity of patients. In addition, activity in the left amygdala was correlated with activity in bilateral posterior parahippocampal gyri, a number of left prefrontal regions, anterior and posterior cingulate, thalamus, and insula. By comparison, controls showed correlations between the left amygdala and fewer regions, mainly in temporal and occipital areas. They interpreted these effects as resulting from patients processing the emotional content of the faces to a greater degree than healthy controls, and that patients used the amygdala in a compensatory role. However, as patients were not likely to have processed the emotional content of the emotionally neutral checkerboard stimuli examined in the present study, this suggests a more general, or less task-dependent, role for the amygdala in brain function of demented patients.

The present results and those of Grady et al (2001) are potentially contradictory when compared with previous research that has found consistent amygdalar volume reduction, hypometabolism and reduced acetylcholine esterase activity in dementia (Braak, Del Tredici and Braak 2003; Callen, Black and Caldwell 2002; Shinotoh et al. 2003). A prediction from these findings might be that the amygdala

would contribute less to the control of global brain function in proportion to decline in its own level of function. However, our results show the opposite, with an increased contribution of the amygdala to the function of other regions in dementia, especially the parietal lobe but also including the average over the entire right cerebrum. A number of hypotheses could be presented to explain this phenomenon. One is that even though mean metabolic activity of demented elderly is lower than in healthy elderly when averaged over minutes or hours (as in typical PET or SPECT studies), the variance in amygdalar activity over short time periods as observed using fMRI (2.6 sec in the present dataset) is relatively higher in demented than healthy elderly. We examined this hypothesis by examining variance in BOLD fMRI signal, and found a larger average magnitude of amygdalar variance in the demented subjects (12.22%) when compared with healthy elderly (8.14%). This did not result from a generally larger variance in the demented subjects overall, as shown by the smaller average variance found for the right cerebrum of demented subjects (1.77%) relative to healthy elderly (2.93%). It is possible that greater variance in amygdalar activity of demented elderly produced a larger effect on global brain function through distal efferents, and so reveal more significant causal networks associated with amygdalar activity than in healthy elderly, even in the presence of reduced volume and metabolism.

Another interesting feature of the networks found in demented patients are that 4 of the best 5 children are in the parietal lobe (BA 7 and 40, inferior parietal lobule and the whole parietal lobe) while none of the best 5 children are in the parietal lobe for the healthy elderly group. In addition to the amygdalar contributions mentioned above, the parents of each of these children in the demented parietal lobe include other parietal regions and portions of the cerebellum and thalamus. It is interesting that these regions all respond to stimuli that require a behavioral response, such as rare target stimuli presented in the oddball task (Clark et al. 2000; 2001; Clark 2002). Indeed, evidence for this target response can be found in the parietal, thalamic and cerebellar regions of subjects in the original analysis of Buckner et al. (2000). Of course, this previous analysis did not take into account causal relationships among brain regions. Even though activity in these regions can be observed for both healthy and demented elderly

groups, the temporal causal relationships among them is more apparent in dementia. The event-related potential response to target stimuli, termed the P300 or P3 response, has been observed even when all stimuli are targets, given that stimuli are spaced widely enough in time (Struber and Polich 2002). The stimuli used by Buckner et al. were spaced approximately 21 seconds apart, which may have been wide enough to elicit a rare target-like response to all stimuli presented.

Differences between demented and healthy control groups could have resulted from a variety of sources. If there were consistent hemodynamic delays between brain regions that differed in a consistent way between healthy and demented elderly, they might confound the causal DBN analysis used here. However, Buckner et al. (2000) found only small hemodynamic differences present between elderly and control subjects, suggesting that this was not the case. Another possible source of differences in the causal networks reported here is that the neurodegeneration associated with dementia altered the location of brain regions sufficiently to disrupt the Talairach normalization procedure. We have examined this by comparing the normalized images within and between groups. We did not find substantial differences between groups to suggest that the normalization procedure failed in the demented group. However, subtle differences could still be present between groups.

**Conclusions**

We have introduced a method for extracting time dependant causal relationship networks of neural anatomical regions based solely on the data contained within a patient's fMRI scan, as well as a method for estimating the confidence in these network structures. In order to address the extremely data intensive nature of fMRI data, we used the Talairach database to group individual voxels into anatomical regions. We modeled the aggregate mean voxel activations of each of the regions with three separate techniques: dynamic Bayesian networks (DBNs), Gaussian naïve Bayesian networks (GNBNs) and support vector machines (SVMs). While the overall classification accuracy of the GNBNs and DBNs were generally equivalent (with both more accurate than the SVM), only the DBNs are capable of

extracting time dependant neural anatomical network structures. Increasing the number of parents each node in the DBNs was shown to increase the overall classification accuracy of the DBNs, which clearly indicates that there is useful temporal information contained within fMRI data.

We noted the importance of quantization on the overall accuracy of the DBNs. While using a specific quantized dataset in performing one type of classification may have yielded good results, use of the same quantized dataset on another classification would frequently yield poor results. As of yet, we have only investigated simple types of quantization and we plan on investigating other quantization schemes in future work. We also plan on testing out continuous DBNs to avoid the issue of quantization all together (albeit at the cost of making a Gaussian distribution assumption—an assumption that is violated given the distribution of AMVA values).

Using this method with visual-motor reaction time task data (Buckner et al. 2000) obtained through the fMRI Data Center, the DBN was able to correctly discriminate demented vs. healthy elderly subjects with 73% accuracy. In addition, the DBN identified causal neural networks present in 93% of healthy elderly studied, including visual-motor networks, and regions involved in emotion, motivation, memory and integrative processing. Greater heterogeneity of causal structure was found in demented elderly, and included consistent input from the amygdala to bilateral parietal cortex as well as the whole right cerebrum, which may be related to anxiety and agitation associated with dementia in its later stages. While we found no explanation for this heterogeneity in the present dataset, it is possible that it may be associated with behavioral, environmental, genetic or other differences that were not examined or reported in the dataset provided. Further research may provide some evidence as to the cause of these within-group differences, and ultimately increase the accuracy of this method.

This is the first publication we are aware of to report a greater correlation of amygdalar activity to brain function in individual demented elderly patients during an emotionally neutral task, and agrees with a previous PET study showing increased correlation between the amygdala and other brain regions across

demented subjects (Grady et al. 2001). Increased contribution of amygdalar control to the function of other brain regions could explain the higher incidence of anxiety and agitation in dementia. This finding suggests that these symptoms might be reduced by disrupting these amygdalar networks.

**Future Work**

Ssignificant issue when dealing with DBNs is that of time delay. In our current work, we only model the interaction between temporally consecutive time frames. Increasing the time delay between examined relationships between neuroanatomical regions drastically increases the size of the search space, however, we are investigation techniques to address these issues. One such technique recently introduced to alleviate the massive search structure problem inherent in DBNs (and BNs in general) is that of Moore and Wong's optimal reinsertion (Moore Wong 2003). They introduce an efficient method for determining the optimal set of parents a node in a BN has and we believe their technique may hold promise in application to our DBNs. A related issue involves the temporal smoothness of the hemodynamic response. Changes in neural metabolic activity are reflected in subsequent changes in blood flow with a significant delay and temporal dispersion or "smearing" (Maisog et al. 1995). This results in significant sharing of information across separate BOLD fMRI acquisitions, and would tend to reduce the magnitude of changes in BOLD response over short periods as examined here using DBNs. While this may have some effect, our results show that the DBNs are still able to identify causal networks even with significant sharing of information across adjacent time points.

The Talairach database allowed us to drastically reduce the amount of data to analyze. Future work includes techniques designed to explicitly model the hierarchical structure of this database. Of particular interest is a statistical technique commonly referred to as shrinkage. Shrinkage allows for the combination of abundant general data with sparse specific data, i.e., it combines low variance/high bias data with high variance/low bias data. This is exactly the relationship of information at different levels of the Talairach hierarchy. At the highest level of granularity, there are only a few number of large

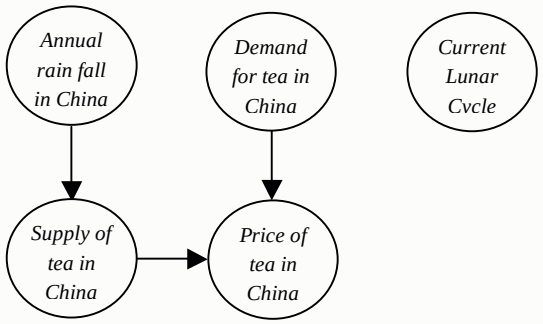
anatomical structures, each based off of copious numbers of voxels, i.e., data with low variance and high bias. At the lowest level in the hierarchy, there are numerous small structures each based off of relatively few voxels, i.e., low bias high variance data. Shrinkage has been successfully applied to other problems in machine learning, including text classification (McCallum, Rosenfeld and Mitchell 1998) and web page classification (Anderson, Domingos and Weld 2002).

The hierarchical nature of the Talairach database also allows for more sophisticated approaches to DBN structure search algorithms in which different levels of granularity of the brain are consecutively searched. A structure search could start with the coarsest level of granularity—which contains a mere six anatomical structures—and perform an exhaustive DBN structure search. The results of that search could be used to guide the structure search for the next level of granularity, which contains too many nodes to perform an exhaustive search on.

The techniques we have applied to the dataset provided by Buckner et al. and future improvements of our technique, may be applied to other fMRI datasets where different patient groups are compared. These methods may eventually prove useful for the diagnosis and treatment of a variety of CNS disorders.

## Methods

### Bayesian and Dynamic Bayesian Networks




a)

| Price | Demand: High | Demand: Low |
|---|---|---|
| High | 0.50 | 0.25 |
| Low | 0.50 | 0.75 |

*Supply = High*

| Price | Demand: High | Demand: Low |
|---|---|---|
| High | 0.95 | 0.35 |
| Low | 0.05 | 0.65 |

*Supply = Low*

b)

Figure 4. a) Example of a hypothetical Bayesian network modeling the price of tea in China. The directed arcs indicate that the price of tea is directly influenced by both the demand and the supply of tea, and indirectly influenced by the annual rain fall in China. Conversely, the absence of an arc indicates that the current lunar cycle has no statistical effect on the price of tea in China. b) The conditional probability table for the *Price of tea in China* node. The table indicates, among other things, that if the supply is *high* and the demand is *low* then 25% of the time the price will be *high* and the 75% of the time the price will be *low.* Note that the *Price* node has two parents resulting in a 3 dimensional conditional probability table. While not shown, there are also tables for all other nodes in the graph.

For an introductory overview of Bayesian networks (BNs), we refer the reader to Charniak's aptly titled "*Bayesian Networks without Tears*" (Charniak 1991), and for a detailed analysis to Heckerman, Geiger and Chickering (1995). In this section, we briefly describe BNs and DBNs in the general case, and in the next section, Dynamic Bayesian Network Analysis, we apply the DBN framework to the fMRI domain.

A Bayesian network is a compact representation of a general joint probability distribution over a set of random variables. It can be thought of as a directed graph with nodes and arcs connecting the nodes. The nodes represent variables in the system being modeled and the arcs represent causal relationship between them.

Figure 4 gives a hypothetical example of a Bayesian network that models the price of tea in China. The arcs in the graph indicate that the price is directly influenced (stochastically) by both the supply and the demand for tea in China. However, the absence of an arc from the *Current Lunar Cycle* to any of the other nodes in the graph indicates that the current lunar cycle does not statistically influence the price of tea in China. The *Price of Tea in China* node is considered a child node with a parent set including the nodes *Demand for tea in China* and the *Supply of tea in China*.

Each node represents a random variable that may take on a set of values. In the example, these values could be {*high* or *low*} for the demand, supply and price of tea in China, {*less than 50 inches, more than 50 inches*} for the annual rain fall and {*new, waxing, full, waning*} for the current lunar cycle. In this network, each node takes on a discrete set of values, however, a Bayesian network can contain continuous-valued nodes as well. Within each of the nodes in a Bayesian network is a set of conditional probabilities. In the discrete case, these probabilities are generally given by a conditional probability table (CPT). The CPT for the *Price of Tea in China* node is shown in Figure 4b. CPTs are collections of multinomial distributions, with a single multinomial describing a node's behavior for every unique setting of the node's parents. In the continuous case, $B_O$ is typically represented by some conditional probability distribution function, such as the conditional Gaussian distribution.

Formally, a Bayesian network is described by the pair $(B_S, B_O)$ where $B_S$ is the graphical structure of the network and $B_O$ specifies the conditional probability information of the BN. $B_S$ contains a set of nodes for every variable in the system, $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$, and an arc for every causal dependency between nodes. In the example given in Figure 4, $B_S$ would include the nodes, *Annual rain fall in China*, *Demand for tea in China*, *Supply of tea in China*, *Price of tea in China* and *Current Lunar Cycle*. It would also contain the three arcs connecting these nodes. $\Pi_i$ indicates the set of nodes that are parents of node $X_i$, and it can take on $q$ possible configurations. For the node *Price of tea in China*, $\Pi_i$ = {*Demand for tea in China, Supply of tea in China*}.

$B_O$ is the set of CPTs for all of the nodes in $B_S$ and encodes the degree and direction to which the nodes in $\Pi_i$ influence $X_i$. If a node has $k$ parents (i.e., $|\Pi_i|=k$), then the CPT for that node will have $k+1$ dimensions specifying $\Pr(X_I \mid \Pi_i)=\Pr(X_I \mid \Pi_i^1, \Pi_i^2, \ldots, \Pi_i^k)$ where $\Pi_i^\ell$ is the $\ell^{th}$ parent of $X_i$. As the number of parents for a node grows, the size of the resulting CPT grows exponentially.

Of particular interest for this work are two specific categories of Bayesian networks: Gaussian naïve Bayesian networks (GNBNs) and dynamic Bayesian networks (DBNs). In a GNBN there are two types of nodes: a single *class* node, and an arbitrary number of *observable* nodes. The conditional probability information within each node is represented as a Gaussian probability density function. A GNBN makes several strong assumptions about the underlying system being modeled: first, the distribution of values in a node are Gaussian distributed and second, the values in each *observed* node are statistically independent of the values in every other *observed* node given the value of the *class* node. These are both strong assumptions that will almost certainly be violated to some degree, though, GNBNs have often been found to perform well in many applications. An example of a 4-state GNBN is given in Figure 5d.

Unlike a GNBN, a DBN explicitly models temporal processes. The DBN is divided into columns of variables where each column represents a time frame in the process being modeled. Each system variable is represented by a single node in each of the columns. Casual relations are allowed to connect nodes between columns, provided the direction of the causal relation indicates a relation from a past time step to a future time step. Isochronal links—links connecting nodes within the same column—may or may not be present in a DBN. An example of an eight state DBN is given in Figure 5g.

Ideally, a DBN would contain one column of variables specifically for every time frame in the system being modeled. However, for all but the simplest models, this is intractable. To reduce the complexity of the DBN model, several assumptions can be (and are frequently) made. The first assumption is the Markov assumption which requires the values a variable takes on at time $t$ can be accurately predicted from only knowledge of its parent variables' values from time $t$-$m$ to time $t$, for some bounded history length $m$[5]. The second assumption is *stationarity* which requires the dynamic behavior of a variable

[5] Frequently, $m$ is set to 1. When it is, the assumption is referred to as a first order Markov assumption.

from time *t* to time *t+1* be independent of *t*. Equation (1) expresses the Markov Property and (2) expresses the *stationarity* assumption,

$$P\left(\boldsymbol{X}^{t} \mid \boldsymbol{X}^{1}, \boldsymbol{X}^{2}, ..., \boldsymbol{X}^{t-1}\right) = P\left(\boldsymbol{X}^{t} \mid \boldsymbol{X}^{t-m}, \boldsymbol{X}^{t-m+1}, ..., \boldsymbol{X}^{t-1}\right), \quad (1)$$

$$P\left(\boldsymbol{X}^{t} \mid \boldsymbol{X}^{t-1}\right) = P\left(\boldsymbol{X}^{t'} \mid \boldsymbol{X}^{t'-1}\right), \forall t, t', \quad (2)$$

where $\boldsymbol{X}^{t}$ is a set of all variables in the column representing time step *t*.

The assumptions made by the DBN are far less restrictive than the assumptions made by the GNBN, and allow for a more sophisticated representation of the data. This causes the DBNs to be significantly more complex than GNBNs and introduces new challenges when DBNs are used as a modeling framework. Most significantly, if the causal structure of the DBN is unknown, as in the case of learning the neuroanatomical networks from fMRI data, learning the DBN structure purely from data is a difficult problem. Technically, DBN structure learning falls into the complexity class *non-deterministic polynomial complete* (NP-Complete) (Chickering, Geiger and Heckerman 1994). The most significant consequence of being NP-Complete is that as the number of nodes within a DBN grows, any process used to find the *optimal* network structure for the DBN must take exponentially longer to complete. To attempt to find good solutions for NP-Complete problems, heuristics for searching the network structure must be devised.

The formulation of the DBN search, and BN search in general, is a well-understood problem (Heckerman 1995). Ideally, every possible structure in the DBN would be individually constructed and the structure that best modeled the data would be chosen. Two significant problems arise. First, a network score indicating how well a given structure fits the data must be devised. Second, as structure search is NP-Complete, it is not feasible to construct every possible structure.

Several scoring metrics have been proposed including AIC (Akike 1973), BIC (Schwarz 1978), BDEU (Buntine 1991), BD (Cooper and Herskovits 1992) and most recently, BDE (Heckerman 1995). We use

Heckerman's BDE score as it is a widely used scoring metric with several significant advantages, such as its ability to easily incorporate prior knowledge and if two structures have the same posterior likelihood given a set of data, the BDE score for the two models will be the same. Like most scoring metrics, the BDE metric is composed of the aggregation of the scores of each of the *families*—a node and the set of its parents—in the network. Equation (3) gives the scoring metric for the entire network, and (4) gives the score for a single *family* in the network,

$$P(D, B_S \mid \xi) \propto \prod_{i=1}^{n} P(X_i, \Pi_i) \cdot score(X_i, \Pi_i, D), \tag{3}$$

$$score(X_i, \Pi_i, D) = \prod_{j=1}^{q_i} \left[ \frac{\Gamma(N'_{ij})}{\Gamma(N_{ij} + N'_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + N'_{ijk})}{\Gamma(N'_{ijk})} \right], \tag{4}$$

where $\rho(X_i,\prod_i)$ is a *structural prior* allowing certain structures to be considered more likely than other structures. $\Gamma(\cdot)$ is the gamma function. $N_{ij}$ is a count for the number of times the parent set of $X_i$ is in configuration $j$. $N_{ijk}$ is the count of the number of times the parent set of $X_i$ is in configuration $j$ while $X_i = k$. $\xi$ refers to the set of prior information, $\left\{ N'_{ij}, N'_{ijk}, \rho(X_t, \Pi_t) \right\} \forall i, j, k$.

Prior information influencing conditional probabilities is allowed in the form of the $N'$ virtual counts, which are analogous to $N$ except they are not counted from dataset $D$ but instead represent some hypothetical prior dataset that describes how variables in the system are thought to behave in relation to each other. For instance, if we wanted to indicate that the *Price of tea in China* should be *low* when the *Demand for tea in China* is *high* and the *Supply of tea in China* is high, we could set the $N'_{ijk}$ count to a positive number where $X_i$ = *Price of tea in China*, $j$ = {*high, high*} and $k$ = *low*. The magnitude of the number would indicate the degree to which we believed this would be the case. Networks connecting the *Price of tea in China* to either the *Supply of tea in China* or the *Demand for tea in China* would then receive a higher score even without sufficient supporting data. In effect, the BDE metric measures the

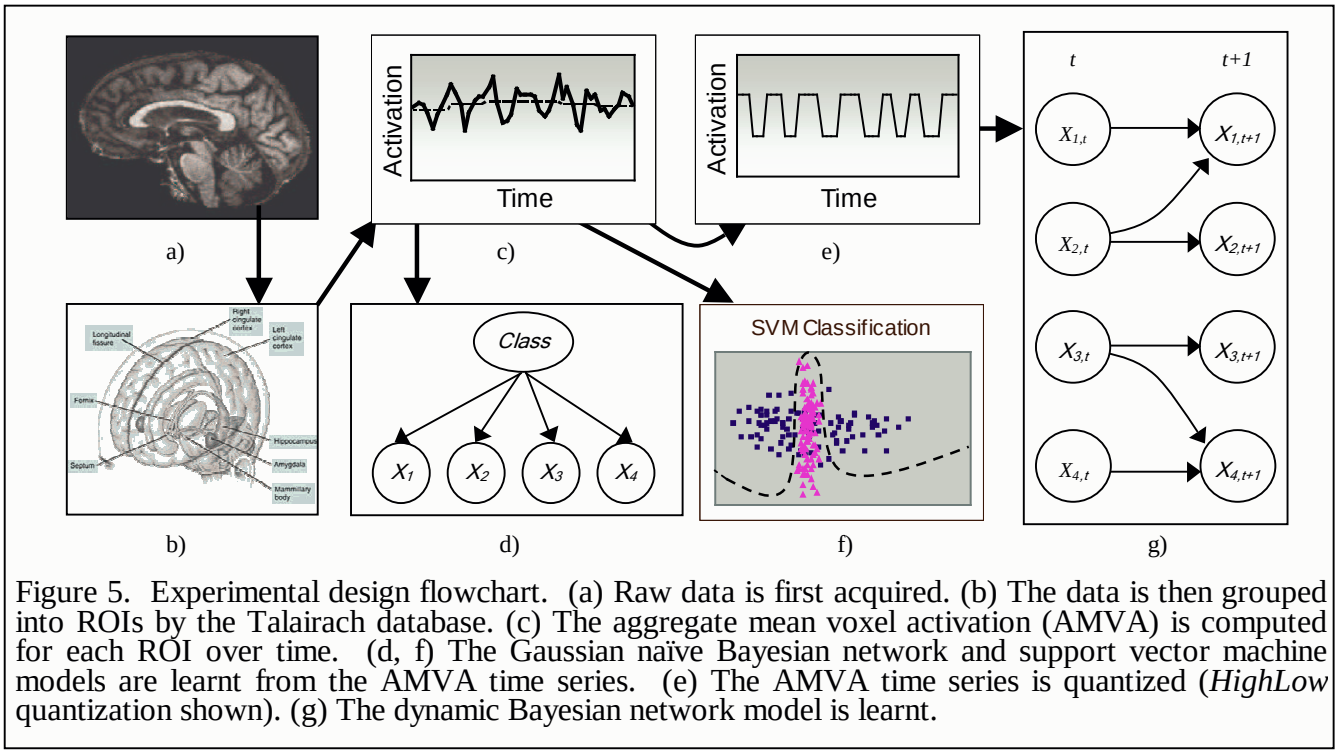



Figure 5. Experimental design flowchart. (a) Raw data is first acquired. (b) The data is then grouped into ROIs by the Talairach database. (c) The aggregate mean voxel activation (AMVA) is computed for each ROI over time. (d, f) The Gaussian naïve Bayesian network and support vector machine models are learnt from the AMVA time series. (e) The AMVA time series is quantized (*HighLow* quantization shown). (g) The dynamic Bayesian network model is learnt.

amount of uniformity in the CPTs' multinomial distributions. The less uniform the distribution, the more informative the parents are and thus, the higher the score will be[6].

**Experimental Design**

Details of the experiment can be found in Buckner et al. (2000). Each patient participated in 60 trials, each of which presented a single or double flickering checkerboard stimulus to the patient. Subjects were instructed to response to the stimulus with a button press. In the single exposure trials, a patient would observe a solitary visual stimulus whereas in the double exposure trials, two temporally separated images would be observed. During each trial, 8 volumes of T2* weighted echo-planar images with a TR of 2.68 sec. were obtained resulting in a time series of 480 fMRI volumes per patient. Each volume contained a grid of 64x64 voxels in 16 slices. This results in a total of 16,384 voxels per image and 7,864,320 voxels per patient.

**Data Preparation**

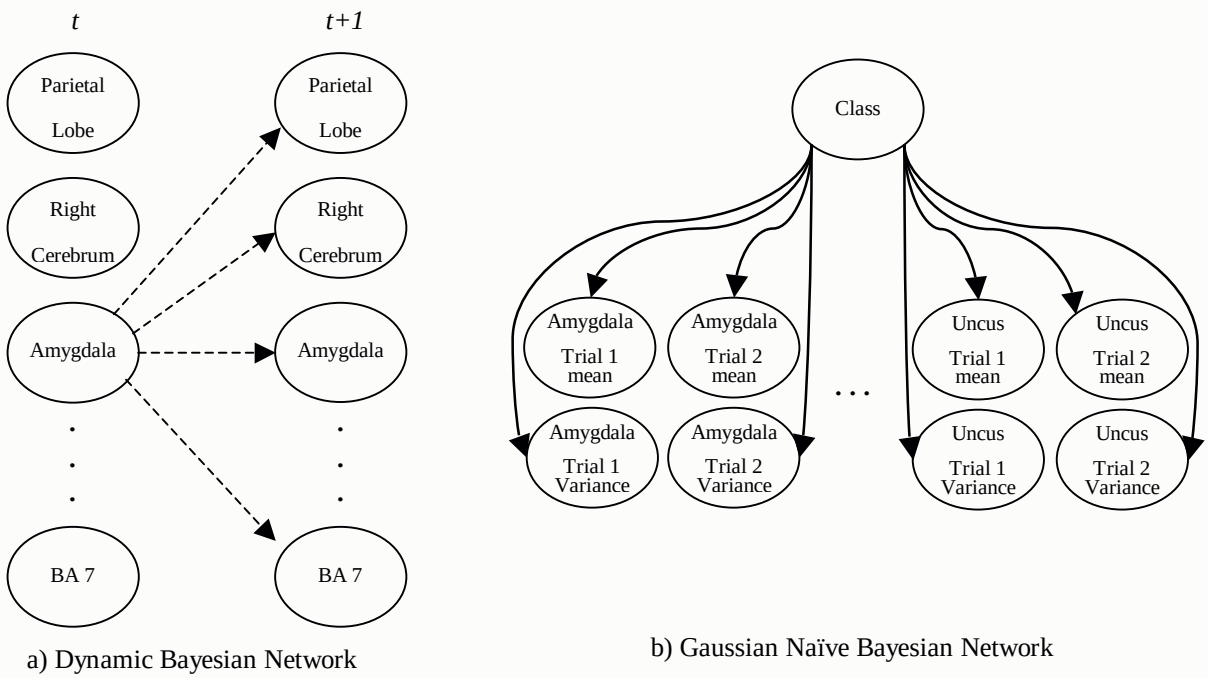
[6] In practice, the network scores are astronomically small, on the order of $10^{-10,000}$, and must be represented in logarithmic space. All results given in this paper for network scores are given as such.

To reduce the enormous quantity of the data produced by an fMRI scan, we abstract from the voxel level to the neuroanatomical region level. Several databases describe such regions of interest (ROIs), with the most widely used being the Talairach database (Lancaster et al. 2000). The Talairach database separates voxels into 150 different regions layered into four hierarchical levels (with further differentiation between gray matter, white matter and cerebro-spinal fluid). Each hierarchical level roughly corresponds to a different level of granularity of the brain, with the topmost level containing the left and right cerebrum, cerebellum and brain stems, and the bottom most level containing specific regions such as the Brodmann areas and individual nuclei. The Talairach data maps every voxel to five different anatomical regions; one for each level in the Talairach hierarchy, and one indicating gray or white matter. We use this mapping to aggregate mean voxel activations to the ROI level and model the relational dependencies between ROIs.

Figure 5 gives a flowchart for the overall experiment. First, raw fMRI data is obtained for each of the patients. Then, for each individual fMRI scans, voxels are grouped into the ROIs they compose (each voxel contributes to 5 regions, one for each level in the Talairach hierarchy). For each structure, the aggregate mean voxel activation (AMVA) is computed resulting in a time series consisting of 480 time points each with 150 attributes. This collection of all the patient's data is referred to as dataset $\boldsymbol{D} = \{D_1, D_2, \ldots, D_{40}\}$ where $D_i$ comprises the data available for the $i^{th}$ patient. Each $D_i$ can be thought of as a table with the columns being Talairach regions, the rows being time slices, and each cell giving the AVMA of a single Talairach region within a single fMRI image.

Four datasets were constructed from $\boldsymbol{D}$: $\boldsymbol{D_h}$, $\boldsymbol{D_d}$, $\boldsymbol{D_e}$ and $\boldsymbol{D_y}$. They each contain data for the healthy patients, demented patients, elderly patients or young patients, respectively. For instance, $\boldsymbol{D_h} = \{D_i \mid \text{class of } D_i = \text{Healthy}\}$. $\boldsymbol{D_e}$ and $\boldsymbol{D_y}$ form a partition of $\boldsymbol{D}$. $\boldsymbol{D_h}$ and $\boldsymbol{D_d}$ form a partition of $\boldsymbol{D_e}$.

At this point, the data is ready to be used by both the SVM and the GNBN. However, the DBN we employ is discrete in nature and requires the data to be quantized. To quantize the data, it is first




a) Dynamic Bayesian Network

b) Gaussian Naïve Bayesian Network

Figure 6. a) A possible structure resulting from the DBN search. The left column represents the values of neuroanatomical regions at time *t* and the right column represents the values of neuroanatomical regions at time *t*+1 which are dependant on regions it time *t*. Arcs are dotted to emphasize that their existence must be searched for, and do not represent a static structure. b) The structure for the GNBN. The time series for each neuroanatomical region is represented by four statistical aggregates; the *Trial1* mean, *Trial2* mean, *Trial1* variance and the *Trial2* variance. The two sets of nodes for the Amygdala and the Uncus are shown, however, a set of four nodes exists in the GNBN for each of the 150 Talairach regions. Unlike the DBN, the GNBN does not require a structure search, and its arcs are represent with solid lines to indicate a static structure.

separated into sequential non overlapping windows of eight fMRI images[7]. A trend is then defined per window as the mean of the AMVA values that compose the window. The offset of the AMVA values from the trend can then be used to tell when the activity of a region is in a 'high' state or a 'low' state.

Three quantized methods are employed**.** The first method results in two quantized states; one state above the trend, and one state below the trend. The second method resulted in four states, two above and below the trend, and the third method resulted in eight states, four above and four below the trend. They were respectively labeled *Highlow*, *Highlow2* and *Highlow4*. Figure 5e. illustrates the *Highlow* quantization for the AMVA time series in Figure 5c. For the *Highlow2* and *Highlow4* quantizations, maximum and minimum AMVA values were determined for each window. This defined an upper and

[7] A window of size eight was chosen since each trial in the experiment resulted in exactly eight fMRI scans.

lower range for quantized states located above and below the trend. The two ranges were then equally divided into 2 or 4 quantization states for the two datasets, respectively.[8]

**Dynamic Bayesian Network Analysis**

Ideally, a DBN modeling the temporal interaction between ROIs in fMRI data would contain *T* columns of *n* nodes, where *T* equals the number of time steps in the experiment (i.e., the number of fMRI images) and *n* equals the number of ROIs. Each node would represent the value of a specific ROI at a specific time step and an arc between the nodes would indicate a temporal relationship between the two ROIs. For example, an arc from node $X_i^{t_1}$ to node $X_j^{t_2}$ would indicate a relationship between the value of the *ith* ROI at time step $t_1$ and the value of the *jth* ROI at time step $t_2$. Unfortunately, learning the structure (where to place the arcs) for such a DBN is intractable. To reduce the size of the model, we make both the first order Markov and *stationary* assumption. This reduces the DBN to two columns, each with *n* nodes. Nodes in each of the columns no longer represent the values of ROIs at absolute time frames, but instead represent the statistical nature of ROIs across all time frames.

Figure 6a illustrates a possible structure for a DBN. In this example, the *Amygdala* node in column *t* is a parent of the *Amygdala, Parietal Lobe, BA 7* and *Right Cerebrum* nodes in column *t+1*. This indicates that there is a statistical relationship between the value of the *Amygdala* at one time step and the values of the child regions in the subsequent time step (and that this statistical relationship exists across all pairs of time steps in the data). As a simple example, assume that the sequence of values the *Amygdala* takes on at time steps 1 through 10 is {*high, high, low, high, high, high, low, low, high, high}* and that the sequence of values the *Parietal Lobe* takes on during time steps 1 through 10 is {*low, high, low, low, low, high, high, low, low, high*}. Notice that whenever the *Amygdala* is *high* at time *t*, there is a 66% chance that the *Parietal Lobe* is *high* at time *t+1* and a 33% chance that it is low. Thus, the value of the

[8] We examined other methods of quantization such as using medians and quartiles as opposed to means and extremum, and creating a separate region surrounding the trend constituting a 'noisy' state, but we found no significant impact on our results.

1. Initialize structure $B_S$ with no arcs connecting nodes in column *t* to nodes in column *t+1*
2. Repeat *numParents* times
3. Repeat for every node, $X_i$, in column *t+1*
4. Repeat for every node, $X_k$, in column *t*
5. Add $X_k$ as a parent of $X_i$, i.e., insert $X_k$ into $\prod_i$
6. Calculate BDE score for $X_i$ (i.e., measure goodness of fit)
7. Remove $X_k$ from $\prod_i$
8. End 4.
9. Permanently add the parent $X_k$ to node $X_i$ with highest BDE score
10. End 3.
11. End 2.
12. Empty the parent set of all but *numBestToKeep* of the highest scoring *families*
13. Return high scoring and high confidence structures

Figure 7. DBN structure search algorithm used to locate structures that describe the data well. Resulting structures contain *families* that describe the relations between neuroanatomical individual ROIs. *numParents* indicates how many parents each node in column *t+1* should have at the end of the algorithm and *numBestToKeep* indicates how many *families* will be allowed to keep their parents. A family is composed of a single node in column *t+1* and all of its parents in column *t*.

*Amygdala* node has some predictive ability on the subsequent value of the *Parietal Lobe*. Given an arc between the two nodes, this statistical relationship would be encoded in the CPT for the *Parietal Lobe* child node.

However, not all arcs in the model are equally meaningful and certain DBN structures will better fit the data. If, for instance, the sequence of values for the *BA 7* node was {high, *low, high, low, low, low, high, high, low, low*}, then the value of *BA 7* would have 100% predictive ability on the subsequent value of the *Parietal Lobe* and a structure with a *BA 7* as a parent of the *Parietal Lobe* would be a better match to the data than a structure with the *Amygdala* as a parent. This would be reflected by the *Parietal Lobe's* CPT being much less uniformly distributed, and would result in a significantly higher BDE score. This simple example only includes pair-wise relationships, but CPTs are capable of modeling much more complex interactions between many simultaneous parents and children nodes. It is the ability of a CPT

to describe complex relationships that sets it apart from continuous models that must make assumptions about the form of distributions, i.e., that they are linear, bi-linear, Gaussian, etc.

Even though the number of possible DBN structures significantly decreases given the Markov and *stationarity* assumptions, there are still far too many structures to search them exhaustively. However, due to the nature of DBN structure search, the parent set for each of the children nodes in the *t+1* column can be searched for independently from the other children nodes. We employ a greedy algorithm, detailed in Figure 7, to perform the structure search. For each node in column *t+1*, the algorithm finds the single best predictive parent in column *t*. It then searches for the second parent that makes the best predictor in conjunction with the previously selected first parent. This process is repeated until a predetermined number of parents has been added. The end result of the searching process is that each ROI is associated with a number of other ROIs that together can predict the its future behavior.

As with all Bayesian techniques, prior information about the system can guide the results of the model. While calculating the BDE score for each proposed network (step 6 of the algorithm), we employ an uninformed structure prior and an uninformed conditional probability prior. This makes the assertion that, prior to seeing any data on the patients, there is no bias towards believing any ROI is more or less likely to appear causally connected in the network, nor is there a bias towards how any discovered relationships should behave. This is achieved by setting $\rho(X_i, \prod_i)=1$, $N'_{ijk}$ to 1 and $N'_{ij} = \sum_{k=1}^{r} N'_{ijk}, \forall i, j, k$, where *r* is the number of states node *k* can be in.

The search algorithm is parameterized by two variables, *numParents* and *numBestToKeep*. *numParents* indicates how many parents each ROI should have. This is required since adding a parent will always result in a higher BDE score but will eventually result in a loss of model generality and lower classification accuracy. *numBestToKeep* indicates how many nodes should be allowed to retain their parents. Generally, classification accuracy improves when only a certain subset of nodes are allowed to

retain their parents. This is because many nodes contribute only noise or over fit the data, and reduce generalization ability of the model.

Using this algorithm, four DBNs were learnt: $DBN_k$, $DBN_d$, $DBN_e$ and $DBN_y$, each obtained by training solely from the $\boldsymbol{D_h}$, $\boldsymbol{D_d}$, $\boldsymbol{D_e}$ and $\boldsymbol{D_y}$ datasets, respectively. In order to classify a new subject's data, the posterior likelihood of each of the models given the subject's data can be calculated. The model with the highest posterior likelihood is then used to determine the class of the new subject. Datasets $\boldsymbol{D_h}$ and $\boldsymbol{D_d}$ were used to differentiate between healthy and demented patients while $\boldsymbol{D_e}$ and $\boldsymbol{D_y}$ were used to differentiate between elderly and young patients.

Tables 1 and 2 give 95% confidence intervals calculated via Equation (5),

$$s = \sqrt{\frac{p(1-p)}{n}}, \quad CI = 1.96 \cdot s \tag{5}$$

where $n$ indicates the number of patients in the dataset, $p$ is the accuracy of the classification and $CI$ is the confidence interval. Given an empirically determined accuracy, and a confidence interval for that accuracy, it is possible to calculate a lower bound that, with 95% probability, the true accuracy is at least equal to or above. With relatively few patients from which to train, the confidence intervals for the DBN and GNBN are approximately 0.13. But, even with this wide confidence interval, with high probability, the two methods are able to classify patients significantly more accurately than would be expected from random chance alone.

**Quantifying DBN Family Network Confidence**

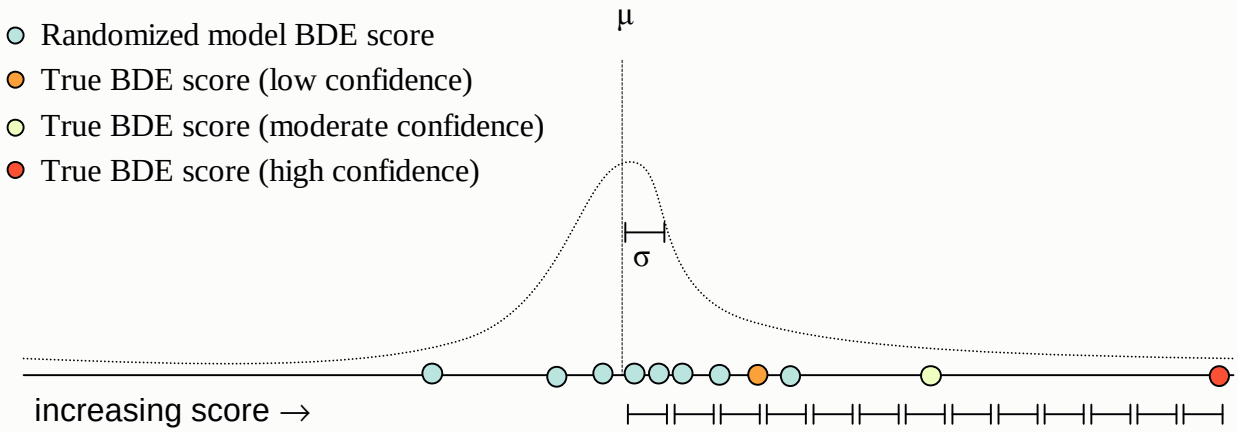



Figure 8. Visualization of the *family* confidence measure. The confidence of *families* within a DBN trained on a true dataset vary in their reliability. Computing the BDE score of *families* in DBNs trained on randomly permuted datasets yields a distribution of BDE scores occurring due to random chance. The larger the distance of a true *family's* BDE score from the BDE random mean, the more likely the true *family* did not occur by chance.

We have developed a technique to measure the confidence in the results returned by the DBN structure search. To measure confidence, we analyze the chance that the predictive ability of a *family* in a network (based on its BDE score) was due to random chance as opposed to an actual underlying temporal relationship among anatomical regions. This is done by training alternative networks on samples from empirical distributions with the same first order statistics, but no temporal correlations.

Recall that the data for each patient is stored in a table where the columns represent the activation values of neuroanatomical regions and the rows represent particular points in time. Normally, the rows are ordered chronologically and dependencies from one time step to the next can be analyzed. We randomly permuted the rows in each of the patient's tables to destroy temporal correlations. For any given row in the table, the relationships between the activation of regions are not affected, however, any temporal relationships among the ROIs are destroyed. The DBN network structure learning algorithm is then applied to the randomly permuted data to calibrate BDE scores for a network based on data with no temporal relationships. This process is repeated 20 times, and a distribution of BDE scores for the randomized data is created for each *family* in the DBN. The BDE scores for the *families* in the networks learnt from the original data can then be compared to the distribution of BDE scores for the *families* in the networks learnt from the randomly permuted data. We can then assign a measure of confidence to

the candidate networks based on the distance between these BDE scores. Equation 6 gives the confidence measure,

$$Conf = \frac{(x - \mu_{rand})}{\sigma_{rand}} \quad (6)$$

where $x$ is the BDE score of the *family* trained on the true dataset, $\mu_{rand}$ is the mean BDE score and $\sigma_{rand}$ is the standard deviation of the BDE score for the equivalent *family* trained on the randomly permuted data (the equivalent *family* trained on the randomly permuted data will have the same child as the original *family)*. Figure 8 gives a visualization of the confidence measure.

For example, within the demented patients, the DBN search reports that the putamen's parent set contains Brodmann area 31, the anterior lobe, the inferior semi-lunar lobule and the lateral dorsal nucleus. The BDE score for this particular *family* is -3359. However, the mean BDE score for the putamen based on the permuted data is -3427 with a standard deviation of 31. The original BDE score of -3359 is only 2.16 standard deviations better than the random BDE mean score. At this distance, there can be little confidence in the reported structure.

Conversely, the DBN network search on the demented data results in the inferior parietal lobule's parent set containing the inferior parietal lobule itself, the uvula of vermis, the amygdala and the inferior occipital gyrus. The BDE score for this family is -3032 (a value significantly higher than the BDE score of -3359 for the putamen's family). The BDE mean for the inferior parietal lobule in the randomized data is -3261 with a standard deviation of 26. The candidate BDE score is 8.52 standard deviations away from the randomized mean BDE score and indicates that it is highly unlikely that a BDE score this high was generated by chance. Table 3 lists the confidences for each of the top 5 network families for both the demented and healthy patients.

The confidence measures for the demented group, while significant, were generally smaller than the confidence measures for the healthy group. This corresponds with the DBN classifier being able to more accurately classify healthy patients than demented patients.

**Support Vector Machine Analysis**

A detailed discussion on support vector machines is outside the scope of this paper. In this section we give the specifics on training the SVMs we employed which would be required to replicate our efforts.

Two datasets are created for analysis with the SVM, $D_{raw}$ and $D_T$. $D_{raw}$ uses the raw fMRI voxel information as dimensions for a single data point. In this dataset, each patient is represented by sixty 65,536 dimensional data points (one data point per fMRI image and 60 data points per patient). $D_T$ contains the AMVA values the DBN used for each of the Talairach regions as dimensions of a single data point resulting in each patient being represented by sixty 150 dimensional data points.

Each dataset is then trained on an SVM with a linear and a Gaussian kernel. As SVMs are sensitive to having an imbalance in the number of representatives in each class, we modify the diagonal elements of the kernel matrices as suggested by Veropoulos, Campbell and Cristianini (1999). We used the Proximal SVM (Fung and Mangasarian 2001) to perform the actual SVM classification.

**Gaussian Naïve Bayesian Network Analysis**

Two separate GNBNs were constructed, $GNBN_h$ and $GNBN_a$. $GNBN_h$ was trained on $\boldsymbol{D_h} \cup \boldsymbol{D_d}$ and was used to classify whether or not an elderly patient was demented. $GNBN_a$ was trained on $\boldsymbol{D_e} \cup \boldsymbol{D_y}$ and was used to classify a patient as either young or elderly.

Each GNBN contained a single binary *class* node and 600 observable nodes. The 600 nodes were broken up into groups of four nodes with each group representing statistics on the entire fMRI data time series for a single one of the 150 anatomical regions. The four nodes in each of the regions were the $Trial1Mean^{(i)}$, $Trial1Var^{(i)}$, $Trial2Mean^{(i)}$ and $Trial2Var^{(i)}$ nodes where ($i$) indicates the *ith* anatomical

region. The Gaussian distributions in each of the nodes were the maximum likelihood estimate (MLE) distributions for either the means or the variances of the AMVA values across all of a patient's fMRI images. The means and variances for the *Trial1* nodes were computed only on fMRI images that were scanned while the patient performed a trial with a single visual stimulus. Similarly, the means and variances for the *Trial2* nodes were computed on images taken when a patient performed a trial with 2 visual stimuli. The structure for the GNBN is given in Figure 6b.

Classification of a patient as either healthy vs. demented or young vs. old was performed via standard BN inference. In one experiment, all 600 nodes were used for classification and in another experiment, only the best 50 nodes were used in the classification. 'Best' in this sense represents the nodes that contained two Gaussian distributions which gave the best class discriminative capabilities. LOOCV accuracies are given in Table 1 and ROC curves are given in Figures 1 and 2.

## Acknowledgements

We would like to acknowledge Dr. Randy Buckner and colleagues, and the Dartmouth fMRI Data Center for providing access the dataset used here. This project was partially funded through grant DA012852 from the National Institute of Drug Abuse, NIH.

## References

Azari, N. P., Pettigrew, K. D., Schapiro, M. B., Haxby J. V., Grady C. L., Pietrini, P., Salerno, J. A., Heston, L. L., Rapoport S. I. And Horwitz (1993), B. Early Detection of Alzheimer's Diesease: A Statistical Approach Using Positron Emission Tomographic Data. *Journal of Cerebral Blood Flow and Metabolism.* v. 13, pg. 438-447.

Akike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory*. B.N. Petrov and F. Csaki (eds.), Academiai Kiado, Budapest, pp 267-281.

Anderson, C., Domingos, P. and Weld, D. (2002) Relational Markov models and their application to adaptive web navigation. *International Conference on Knowledge Discovery and Data Mining* (KDD). 143-152.

Braak, H., Del Tredici, K. and Braak, E. (2003). Spectrum of pathology. In: R.C. Petersen, (Ed); *Mild cognitive impairment: Aging to Alzheimer's disease.* London: Oxford University Press. pp. 149-189.

Buckner, R. L., Snyder, A., Sanders, A., Marcus, R., Morris, J. (2000). Functional Brain Imaging of Young, Nondemented, and Demented Older Adults. *Journal of Cognitive Neuroscience*, 12, 2. 24-34.

Buntine, W. (1991). Theory Refinement on Bayesian Networks. *Proceedings of the Seventh Conference on UAI*. 52-60.

Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Kluwer Academic Publishers.

Charniak, E. (1991). Bayesian Networks Without Tears. *AI Magazine*, 12 (4).

Chickering, D., Geiger, D., Heckerman, D. (1994). Learning Bayesian Networks is NP-Hard (Technical Report MSR-TR-94-17). Microsoft.

Clark, V.P. (2002) Orthogonal polynomial regression for the detection of response variability in event-related fMRI. *NeuroImage*, 17, 344-363.

Clark, V. P., Fannon, S., Lai, S., Benson, R. (2001). Paradigm-dependent modulation of event-related fMRI activity evoked by the oddball task. *Human Brain Mapping*, *14*(2), 116-127.

Clark, V. P., Fannon, S., Lai, S., Benson, R., Bauer, L. (2000). Responses to rare visual target and distractor stimuli using event-related fMRI. *Journal of Neurophysiology*, *83*(5), 3133-3139.

Cooper, G., Ninio, M., Friedman, N. and Schuurmans, D. (2002). Data Perturbation for Escaping Local Maxima in Learning. *Proceedings of AAAI-02*. 132-139.

fMRIDC (2004). The fMRI Data Center. Dartmouth. Available at http://lx50.fmridc.org/f/fmridc.

Friston, J. (2002). Statistical Parametric Mapping. Available at http://www.fil.ion.ucl.ac.uk/spm/papers/SPM-Chapter.pdf.

Friston, J. et al. (1995). Statistical Parametric Maps in Functional Imaging: A General Linear Approach. *Human Brain Mapping* 2, 189-210.

Friston, K.J., Harrison, L. and Penny, W. (2003) Dynamic causal modeling. *NeuroImage* 19, 1273–1302.

Goutte, C., Toft, P., Rostrup E., Nielsen, F. and Hansen, L. (1999) On Clustering fMRI Time Series. *NeuroImage* 9, 298-310.

Heckerman, D., Geiger, D., Chickering, D.M. (1995). Learning Bayesian Networks: the Combination of Knowledge and Statistical Data. *Machine Learning*, 20, 197-243.

H∅jen-S∅rensen, P., Hansen, L., Rasmussen, C. (2000). Bayesian Modelling of fMRI Time Series. In S.A. Solla, T.K. Leen and K. R. Muller (eds.), *Advances in Neural Information Processing Systems*. MIT Press.

Lancaster JL, Woldorff MG, Parsons LM, Liotti M, Freitas CS, Rainey L, Kochunov PV, Nickerson D, Mikiten SA, Fox PT. (2000). Automated Talairach Atlas labels for functional brain mapping. *Human Brain Mapping* 10,120-131.

Fung, G., Mangasarian, O. L. (2001). Proximal Support Vector Machine Classifiers. *Proceedings of the seventh ACM SIGKDD*. San Francisco. pp. 77-86.

Maisog, J.Ma., Clark, V.P., Courtney, S., Haxby, J.V. (1995). Estimating the hemodynamic response and effective degrees of freedom in functional MRI time series. *Human Brain Mapping*, Supplement 1:147.

McCallum, A., Rosenfeld, R., Mitchell, T. and Ng, A. Y. (1998) Improving Text Classification by Shrinkage in a Hierarchy of Classes. *International Conference on Machine Learning*.

Mitchell, T., Hutchinson, R., Just, M., Niculescu, R.S., Pereira, F., Wang, X. (2003). Classifying Instantaneous Cognitive States from fMRI Data. *American Medical Informatics Association Annual Symposium*).

Moore, A., Wong, W. (2003) Optimal Reinsertion: A New Search Operator for Accelerated and More Accurate Bayesian Network Structure Search Learning. *Proceedings of the Twentieth International Conference on Machine Learning*.552-559.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Veropoulos, K., Campbell, C., Cristianini, N. (1999). Controlling the sensitivity of support vector machines. In *Proceedings of the International Joint Conference on Artificial Intelligence, Workshop ML3, 55-60*. Stockholm, Sweden.

Volicer L, Hurley AC. (2003) Management of behavioral symptoms in progressive degenerative dementias. *Journals of Gerontology Series A Biological Sciences and Medical Sciences.* 58:M837-845.

Worsley, K. and Friston, K. (1995). Analysis of fMRI times-series revisited – again. *Neuroimage*, 2, 173-181.