Neural classifiers for schizophrenia diagnostic support on diffusion imaging data

Alexandre Savio, Juliette Charpentier, Maite Termenón, Ann K. Shinn, Manuel Graña *

Abstract: Diagnostic support for psychiatric disorders is a very interesting goal because of the lack of biological markers with sufficient sensitivity and specificity in psychiatry. The approach consists of a feature extraction process based on the results of Pearson correlation of known measures of white matter integrity obtained from diffusion weighted images: fractional anisotropy (FA) and mean diffusivity (MD), followed by a classification step performed by statistical support vector machines (SVM), different implementations of artificial neural networks (ANN) and learn vector quantization (LVQ) classifiers. The most discriminant voxels were found in frontal and temporal white matter. A total of 100% classification accuracy was achieved in almost every case, although the features extracted from the FA data yielded the best results. The study has been performed on publicly available diffusion weighted images of 20 male subjects.

Key words: *DWI, Schizofrenia, Neural Classifiers, Fractional Anisotropy, Mean Diffusivity*

1. Introduction

There is growing research effort devoted to the development of automated diagnostic support tools that may help clinicians perform their work with greater accuracy and efficiency. In medicine, diseases are often diagnosed with the aid of biological markers, including laboratory tests and radiologic imaging. The process of diagnosis becomes more difficult, however, when dealing with psychiatric disorders, in which diagnosis relies primarily on the patient's self-report of symptoms and the presence or absence of characteristic behavioral signs. Schizophrenia is a disabling psychiatric disorder characterized by hallucinations, delusions, disordered thought/speech, disorganized behavior, emotional withdrawal, and functional decline [2]. Currently, diagnosis is made almost exclusively on subjective measures like self-report, observation, and clinical history.

A large number of magnetic resonance imaging (MRI) morphological studies have shown subtle brain abnormalities to be present in schizophrenia. Structural

*A. Savio, M. Termenón, M. Graña are with the Grupo de Inteligencia Computacional (GIC), Universidad del País Vasco, Spain
J. Charpentier is with the Institut Superieur de BioSciences de Paris (ISBS), ESIEE, Université Paris-Est, France
A.K. Shinn is with the McLean Hospital, Belmont, Massachusetts; Harvard Medical School, Boston, Massachusetts, US

studies have found enlargement of the lateral ventricles, particularly the temporal horn of the lateral ventricles [28];[28]; reduced volumes of medial temporal structures (hippocampus, amygdala, and parahippocampal gyrus) [4, 17, 29], superior temporal gyrus [17], prefrontal cortex [15, 32], and inferior parietal lobule[27, 14]; and reversal of normal left greater than right volume in male patients with schizophrenia [24, 12]. In 1984, Wernicke [35] proposed that schizophrenia might involve altered connectivity of distributed brain networks that are diverse in function and that work in concert to support various cognitive abilities and their constituent operations. Consistent with the "dysconnectivity hypothesis", studies have found correlations between prefrontal and temporal lobe volumes [36, 7] and disruptions of functional connectivity between frontal and temporal lobes in schizophrenia [23]. These findings strongly point to widespread problems of connectivity in schizophrenia.

Diffusion tensor imaging (DTI) is a MRI method that allows more direct investigation into the integrity of white matter (WM) fibers, and thus into the anatomical connectivity of different brain regions. DTI depends upon the motion of water molecules to provide structural information in vivo [25, 5], and yields measures like fractional anisotropy (FA) and mean diffusivity (MD). The most commonly demonstrated DTI abnormalities in schizophrenia are decreased FA in the uncinate fasciculus (a tract connecting temporal and frontal regions and involved in decision-making, emotions, and episodic memory), the cingulum bundle (a tract interconnecting limbic regions which involved in attention, emotions, and memory), and the arcuate fasciulus (a tract connecting language regions) [21]. Lower anisotropic diffusion within white matter may reflect loss of coherence of WM fiber tracts, to changes in the number and/or density of interconnecting fiber tracts, or to changes in myelination [19, 22, 1, 20].

The present paper will focus on the application of machine learning (ML) algorithms for the computer aided diagnosis (CAD) of schizophrenia, on the basis of feature vectors extracted from DTI measures of WM integrity, FA and MD. This feature extraction method is based on Pearson correlation, and is simpler than others found in the literature [13, 11]. These features will be the input for statistical SVM and artificial neural networks (ANN) classifiers. We found literature on the application of ML algorithms to the discrimination of schizophrenia patients from healthy subjects. A minimum recognition error of 17,8% using geometry features and FA of DTI from a database of 36 healthy subjects and 34 patients with schizophrenia was reported in [34]. A study of the effect of principal component analysis (PCA) and discriminant PCA (DPCA) was carried on FA volumes reaching a minimum one-leave-out validation classification error 20% using Fisher linear discriminant (FLD) in [9]. Good classification results were also obtained in structural MRI (sMRI) studies [37, 11].

Section 2. gives a summary of the classification algorithms used for this study. Section 3. describes the materials and methods in the study: characteristics of the subjects conforming the database for the study, the acquisition protocol, the preprocessing steps of the MRI and DTI volumes and the feature extraction process. Section 4. gives the results of our computational experiments. Section 5. gives our final comments and conclusions.

2. Neural Network and Statistical Classification Algorithms

We deal with two class classification problems, given a collection of training/testing input feature vectors $X = \{\mathbf{x}_i \in \mathbb{R}^n,\ i = 1, \ldots, l\}$ and the corresponding labels $\{y_i \in \{-1, 1\},\ i = 1, \ldots, l\}$, which sometimes can be better denoted in aggregated form as a binary vector $\mathbf{y} \in \{-1, 1\}^l$. The algorithms described below build some classifier systems based on this data. The simplest algorithm is the 1-nearest neighbor (1-NN) which involves no adaptation and uses all the training data samples. The classification rule is of the form:

$$c\left(\mathbf{x}\right) = y_{i^*} \text{ where } i^* = \arg \min_{i=1,\ldots,l} \left\{\|\mathbf{x} - \mathbf{x}_i\|\right\},$$

that is, the assigned class is that of the closest training vector. To validate their generalization power we use ten-fold cross-validation.

2.1 Support Vector Machines

The support vector machine (SVM) [33] approach to build a classifier system from the given data consists in solving the following optimization problem:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i, \tag{1}$$

subject to

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq (1 - \xi_i),\ \xi_i \geq 0,\ i = 1, 2, \ldots, n. \tag{2}$$

The minimization problem is solved via its dual optimization problem:

$$\min_{\alpha} \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha}, \tag{3}$$

subject to

$$\mathbf{y}^T\boldsymbol{\alpha} = 0,\ 0 \leq \alpha_i \leq C,\ i = 1, \ldots, l. \tag{4}$$

Where $\mathbf{e}$ is the vector of all ones, $C > 0$ is the upper bound on the error, Q is an $l \times l$ positive semidefinite matrix, whose elements are given by the following expression:

$$Q_{ij} \equiv y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{5}$$

where

$$K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j), \tag{6}$$

is the kernel function that describes the behavior of the support vectors. Here, training vectors $\mathbf{x}_i$ are mapped into a higher (maybe infinite) dimensional space by the function $\phi(\mathbf{x}_i)$. The decision function is:

$$sgn(\sum_{i=1}^{l} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b). \tag{7}$$

The regularization parameter C is used to balance the model complexity and the training error. It was always set to 1 in this case study.

The chosen kernel function results in different kinds of SVM with different performance levels, and the choice of the appropriate kernel for a specific application is a difficult task. In this study we only needed to use a linear kernel, defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = 1 + \mathbf{x}_i^T \mathbf{x}_j, \tag{8}$$

this kernel shows good performance for linearly separable data.

2.2 Backpropagation

Backward propagation of errors, or backpropagation (BP), [26, 16] is a non-linear generalization of the squared error gradient descent learning rule for updating the weights of artificial neurons in a single-layer perceptron, generalized to feed-forward networks, also called multi-Layer perceptron (MLP). Backpropagation requires that the activation function used by the artificial neurons (or "nodes") is differentiable with its derivative being a simple function of itself. The backpropagation of the error allows to compute the gradient of the error function relative to the hidden units. It is analytically derived using the chain rule of calculus. During on-line learning, the weights of the network are updated at each input data item presentation. We have used the resilient backpropagation, which uses only the derivative sign to perform the weight updating.

We restrict our presentation of BP to train the weights of the MLP for the current two class problem. Let the instantaneous error E_p be defined as:

$$E_p(\mathbf{w}) = \frac{1}{2}\left(y_p - z_K(\mathbf{x}_p)\right)^2, \tag{9}$$

where y_p is the p-th desired output y_p, and $z_K(\mathbf{x}_p)$ is the network output when the p-th training exemplar $\mathbf{x}_p$ is inputted to the MLP composed of K layers, whose weights are aggregated in the vector $\mathbf{w}$. The output of the j-th node in layer k is given by:

$$z_{k,j}(\mathbf{x}_p) = f\left(\sum_{i=0}^{N_{k-1}} w_{k,j,i} z_{k-1,i}(\mathbf{x}_p)\right), \tag{10}$$

where $z_{k,j}$ is the output of node j in layer k, N_k is the number of nodes in layer k, $w_{k,j,i}$ is the weight which connects the i-th node in layer $k-1$ to the j-th node in layer k, and $f(\cdot)$ is the sigmoid nonlinear function, which has a simple derivative:

$$f'(\alpha) = \frac{df(\alpha)}{d\alpha} = f(\alpha)\left(1 - f(\alpha)\right). \tag{11}$$

The convention is that $z_{0,j}(\mathbf{x}_p) = \mathbf{x}_{p,j}$. Let the total error E_T be defined as follows:

$$E_T(\mathbf{w}) = \sum_{p=1}^{l} E_p(\mathbf{w}), \tag{12}$$

where l is the cardinality of X. Note that E_T is a function of both the training set and the weights in the network. The backpropagation learning rule is defined as follows:

$$\Delta w(t) = -\eta \frac{\partial E_p(\mathbf{w})}{\partial w} + \alpha \Delta w(t-1), \tag{13}$$

where $0 < \eta < 1$, which is the learning rate, the momentum factor α is also a small positive number, and w represents any single weight in the network. In the above equation, $\Delta w(t)$ is the change in the weight computed at time t. The momentum term is sometimes used ($\alpha \neq 0$) to improve the smooth convergence of the algorithm. The algorithm defined by equation (13) is often termed as *instantaneous backpropagation* because it computes the gradient based on a single training vector. Another variation is *batch backpropagation*, which computes the weight update using the gradient based on the total error E_T.

To implement this algorithm we must give an expression for the partial derivative of E_p with respect to each weight in the network. For an arbitrary weight in layer k this can be written using the Chain Rule:

$$\frac{\partial E_p(\mathbf{w})}{\partial w_{k,j,j}} = \frac{\partial E_p(\mathbf{w})}{\partial z_{k,j}(\mathbf{x}_p)} \frac{\partial z_{k,j}(\mathbf{x}_p)}{\partial w_{k,j,i}}. \tag{14}$$

Because the derivative of the activation function follows equation 11, we get:

$$\frac{\partial z_{k,j}(\mathbf{x}_p)}{\partial w_{k,j,i}} = z_{k,j}(\mathbf{x}_p)\left(1 - z_{k,j}(\mathbf{x}_p)\right) z_{k-1,j}(\mathbf{x}_p), \tag{15}$$

and

$$\frac{\partial E_p(\mathbf{w})}{\partial z_{k,j}(\mathbf{x}_p)} = \sum_{m=1}^{N_{k+1}} \frac{\partial E_p(\mathbf{w})}{\partial z_{k+1,m}(\mathbf{x}_p)} z_{k+1,m}(\mathbf{x}_p)\left(1 - z_{k+1,m}(\mathbf{x}_p)\right) w_{k+1,m,j},$$

which at the output layer corresponds to the output error :

$$\frac{\partial E_p(\mathbf{w})}{\partial z_K(\mathbf{x}_p)} = z_L(\mathbf{x}_p) - y_p. \tag{16}$$

2.3 Radial Basis Function Networks

Radial basis function networks (RBF) [10] are a type of ANN that use radial basis functions as activation functions. RBFs consist of a two layer neural network, where each hidden unit implements a radial activated function. The output units compute a weighted sum of hidden unit outputs. Training consists of the unsupervised training of the hidden units followed by the supervised training of the output units weights. RBFs have their origin in the solution of a multivariate interpolation

problem [8]. Arbitrary function $g(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ can be approximated by a map defined by a RBF network with a single hidden layer of K units:

$$\hat{g}_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j=1}^{K} w_j \phi(\sigma_j, \|\mathbf{x} - \mathbf{c}_j\|), \tag{17}$$

where $\boldsymbol{\theta}$ is the vector of RBF parameters including $w_j, \sigma_j \in \mathbb{R}$, and $\mathbf{c}_j \in \mathbb{R}^n$; let us denote $\mathbf{w} = (w_1, w_2, \ldots, w_p)^T$, then the vector of RBF parameters can be expressed as $\boldsymbol{\theta}^T = (\mathbf{w}^T, \sigma_1, \mathbf{c}_1^T, \ldots, \sigma_K, \mathbf{c}_K^T)$. Each RBF is defined by its center $\mathbf{c}_j \in \mathbb{R}^n$ and width $\sigma_j \in \mathbb{R}$, and the contribution of each RBF to the network output is weighted by w_j. The RBF function $\phi(\cdot)$ is a nonlinear function that monotonically decreases as $\mathbf{x}$ moves away from its center $\mathbf{c}_j$. The most common RBF used is the isotropic Gaussian:

$$\hat{g}_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j=1}^{p} w_j \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_j\|^2}{2\sigma_j^2}\right).$$

The network can be thought as the composition of two functions $\hat{g}_{\boldsymbol{\theta}}(\mathbf{x}) = W \circ \Phi(\mathbf{x})$, the first one implemented by the RBF units $\Phi : \mathbb{R}^n \to \mathbb{R}^K$ performs a data space transformation which can be a dimensionality reduction or not, depending on whether $K > n$. The second function corresponds to a single layer linear Perceptron $W : \mathbb{R}^K \to \mathbb{R}$ giving the map of the RBF transformed data into the class labels. Training is accordingly decomposed into two phases. First a clustering algorithm is used to estimate the Gaussian RBF parameters (centers and variances). Afterwards, linear supervised training is used to estimate the weights from the hidden RBF to the output. In order to obtain a binary class label output, a hard limiter function is applied to the continuous output of the RBF network.

2.4 Probabilistic Neural Networks

A probabilistic neural network (PNN) [31] uses a kernel-based approximation to form an estimate of the probability density function of categories in a classification problem. In fact, it is a generalization of the Parzen windows distribution estimation, and a filtered version of the 1-NN classifier. The distance of the input feature vector $\mathbf{x}$ to the stored patterns is filtered by a RBF function. Let us denote the data sample partition as $X = X_1 \cup X_{-1}$, where $X_1 = \{\mathbf{x}_1^1, \ldots, \mathbf{x}_{n_1}^1\}$ and $X_1 = \left\{\mathbf{x}_1^{-1}, \ldots, \mathbf{x}_{n_{-1}}^{-1}\right\}$. That is, superscripts denote the class of the feature vector and $n_1 + n_{-1} = n$. Each pattern $\mathbf{x}_j^i$ of training data sample is interpreted as the weight of the j-th neuron of the i-th class. Therefore the response of the neuron is computed as the probability of the input feature vector according to a Normal distribution centered at the stored pattern:

$$\Phi_{i,j}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{\left\|\mathbf{x} - \mathbf{x}_j^i\right\|}{2\sigma^2}\right]. \tag{18}$$

Therefore the output of the neuron is inside $[0, 1]$. The tuning of a PNN network depends on selecting the optimal sigma value of the spread σ of the RBF functions,

which can be different for each class. In this paper an exhaustive search for the optimal spread value in the range (0, 1) for each training set has been done. The output of the PNN is an estimation of the likelihood of the input pattern $\mathbf{x}$ being from class $i \in \{-1, 1\}$ by averaging the output of all neurons that belong to the same class:

$$p_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} \Phi_{i,j}(\mathbf{x}). \tag{19}$$

The decision rule based on the output of all the output layer neurons is simply:

$$\hat{y}(\mathbf{x}) = \arg\max_i \{p_i(\mathbf{x})\}, \quad i \in \{-1, 1\}, \tag{20}$$

where $\hat{y}(\mathbf{x})$ denotes the estimated class of the pattern $\mathbf{x}$. If the a priori probabilities for each class are the same, and the losses associated with making an incorrect decision for each class are the same, the decision layer unit classifies the pattern $\mathbf{x}$ in accordance with the optimal Bayes' rule.

2.5 Learning Vector Quantization Neural Network

Learning vector quantization (LVQ), as introduced by Kohonen [18], represents every class $c \in \{-1, 1\}$ by a set $W(c) = \{\mathbf{w}_i \in \mathbb{R}^n; i = 1, \ldots, N_c\}$ of weight vectors (prototypes) which tessellate the input feature space. Let us denote W the union of all prototypes, regardless of class. If we denote c_i the class the weight vector $\mathbf{w}_i \in W$ is associated with, the decision rule that classifies a feature vector $\mathbf{x}$ is as follows:

$$c(\mathbf{x}) = c_{i^*}$$

where

$$i^* = \arg\min_i \{\|\mathbf{x} - \mathbf{w}_i\|\}.$$

The training algorithm of LVQ aims at minimizing the classification error on the given training set, i.e., $E = \sum_j (y_j - c(\mathbf{x}_j))^2$, modifying the weight vectors on the presentation of input feature vectors. The heuristic weight updating rule is as follows:

$$\Delta\mathbf{w}_{i^*} = \begin{cases} \epsilon \cdot (\mathbf{x}_j - \mathbf{w}_{i^*}) & \text{if } c_{i^*} = y_j \\ -\epsilon \cdot (\mathbf{x}_j - \mathbf{w}_{i^*}) & \text{otherwise} \end{cases}, \tag{21}$$

that is, the input's closest weight is adapted either toward the input if their classes match, or away from it if not. This rule is highly unstable, therefore, the practical approach consists in performing an initial clustering of each class data samples to obtain an initial weight configuration using equation 21 to perform the fine tuning of the classification boundaries. This equation corresponds to a LVQ1 approach. The LVQ2 approach involves determining the two input vector's closest weights. They are moved toward or away the input according to the matching of their classes.

3. Materials and Methods

Structural MRI and DTI data from twenty men (aged 21-55 yr), ten patients and ten controls, from a publicly available database from the National Alliance for Medical Image Computing (NAMIC) [1] were the subjects of this study in this experiment. The imaging parameters and demographic information about the subjects can be obtained from the web site, we omit them for lack of space. A technical description of the feature extraction method and the data will be available [2], because many of the difficulties found have no place in an academic paper, but are important for the reproducibility of the results.

3.1 Scalar Features of Diffusion Tensors

In DTI, a diffusion tensor at a voxel is a 3×3 positive-definite symmetric matrix D, which can be represented by its decomposition as $D = \lambda_1 \mathbf{g}_1 \mathbf{g}_1^T + \lambda_2 \mathbf{g}_2 \mathbf{g}_2^T + \lambda_3 \mathbf{g}_3 \mathbf{g}_3^T$, where $\lambda_1 \geq \lambda_2 \geq \lambda_3$ and $\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$ are the eigenvalues and eigenvectors of D, respectively. Two scalar measures were extracted [6] from the voxels diffusion tensors: the mean diffusivity (MD) and the fractional anisotropy (FA). The first corresponds to the average eigenvalue:

$$MD = \frac{\mathrm{Tr}\,(D)}{3} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3}. \tag{22}$$

The FA measures the fraction of the magnitude of D that can be related to anisotropic diffusion in a mean-squared sense (i.e. the extent of deviation from isotropic diffusivity in all direction). Its magnitude is also rotationally invariant, and independent from sorting of the eigenvalues. The FA is calculated as follows:

$$FA = \sqrt{\frac{1}{2}} \frac{\sqrt{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}. \tag{23}$$

Thus, isotropic diffusion is imaged as zero value and FA maximum value is one. Figure 1 show slices of FA and MD volumes of one study subject.

3.2 Image preprocessing

Feature extraction requires that the diffusion related data is spatially normalized, in order to compute the correlation measure and to extract the values of the feature vectors. Our starting point was the nonlinear registration [3] of the T1-weighted sMRI skull stripped volumes of each subject to the Montreal Neurological Institute (MNI152) standard template, using the ANTS[3] nonlinear elastic registration algorithm. For the elastic registration, a probabilistic correlation similarity metric was chosen with window radius 4 and gradient step length 1. The optimization has been performed over three resolutions with a maximum of 100 iterations at the coarsest level, 100 at the next coarsest and 10 at the full resolution. The optimization stops when either the distance between both images cannot be further minimized or the

[1] http://www.insight-journal.org/midas/collection/view/190?path_navigation=17
[2] http://www.ehu.es/ccwintco/index.php/GIC-experimental-databases
[3] http://www.picsl.upenn.edu/ANTS

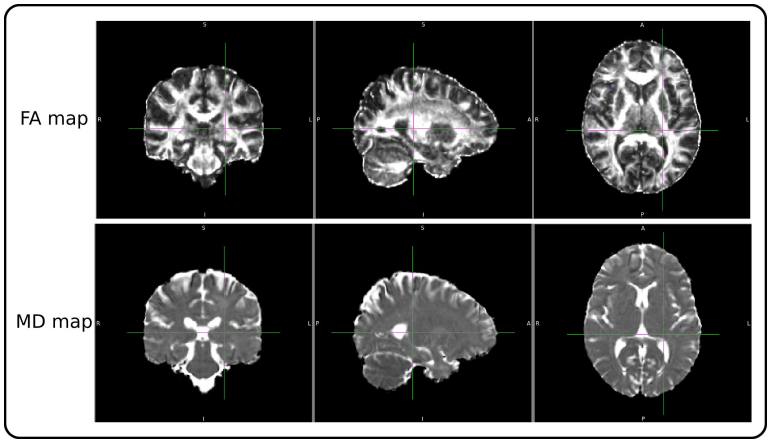

Figure 1 *FA and MD maps of one subject.*

maximum number of iterations is reached. We used a Gaussian regularization with sigma parameter value 3 which operates only on the deformation field and not on the similarity gradient. In addition, a previous histogram matching step has been performed. The deformation fields of this registration were used afterwards for the spatial deformation of the FA and MD volumes.

The DWI scans were already noise filtered and corrected for eddy currents and head motion by the group that originally acquired the scans. A brain mask was obtained for each DWI data volume to calculate the FA and MD maps of each subject [6]. The FA and MD maps were linearly registered to the sMRI skull stripped volumes [30] of each subject and then non-linearly registered to MNI applying the deformation fields obtained from the sMRI data nonlinear registration. All of the FA and MD volumes were then considered spatially normalized.

3.3 Feature extraction

Once the FA and MD maps were spatially normalized, we processed them independently. We considered each voxel site independently, forming a vector at the voxel site across all the subjects. Then we computed the Pearson correlation coefficient between this vector and the control variable with the labels (patient=1, control=−1). Thus we obtained for FA and MD data two independent volumes containing correlation values at each voxel. For each volume we estimated the empirical distribution of the absolute correlation values and determined a selection threshold corresponding to a percentile of this absolute correlation distribution. Voxel sites with absolute value of the correlation above this threshold were retained, and the feature vector for each subject was composed of the FA or MD values at these voxel sites. In table I we show the percentiles and the number of voxels selected for each feature vector.

Although the voxel sites selected to build the feature vectors (the feature mask) were localized in many different regions of the subject brains, we found that most were concentrated in regions of characteristic abnormalities found for schizophrenia

Database	Percentile	DT Measure	Number of voxels
A	99.990%	FA	241
		MD	241
B	99.992%	FA	193
		MD	193
C	99.995%	FA	121
		MD	121
D	99.997%	FA	72
		MD	72
E	99.999%	FA	24
		MD	24

Table I *Databases considered, percentile on the correlation distribution and size of the feature vectors.*

shown in the literature (see [19] for references). The features voxel locations [4] were different for FA and MD maps. In the case of FA, the selected voxels were localized mainly in parietal and temporal lobes, but also in the cerebellum and occipital lobe. More specifically, in WM we found discriminant voxel values in the cingulum bundle, superior and inferior longitudinal fasciculus and in the inferior fronto-occipital fasciculus. On the other hand, in the MD maps, the most discriminant voxel values were the ones localized in frontal and parietal lobes, more specifically thecingulum bundle, inferior fronto-occipital and longitudinal fasciculus, and superior longitudinal fasciculus.

3.4 Classifiers parameters

All classifiers were calculated with a maximum iteration number (epochs) of 100. For the 1-NN classifier, we used the nearest neighbor rule with euclidean distance. In the SVM algorithm, a linear kernel function was used as well as a sequential minimal optimization for the separating hyperplane method. For BPNN, the number of neurons in the hidden layer was 4, the learning rate was set to 0.05, tan-sigmoid transfer function, and training and learning functions were gradient descent with momentum. LVQ2 was trained with 2 hidden neurons, learning rate set to 0.01. The training function used for RBF was according to resilient backpropagation algorithm. In the case of PNN, random order incremental training was used. For the last three algorithms (BPNN, LVQ2 and RBF) zeros were set as initial input and layer delay conditions. These parameters have been selected after a sensitivity analysis.

We tested several cross-validation strategies, because the small database size may have an influence on the results obtained with each of these cross-validation processes. Cross-validation partitions were computed 40 times and we show average accuracy, sensitivity, and specificity for the 10-fold cross-validation procedure.

[4] This specification of the voxel locations were obtained with the "atlasquery" tool from FMRIB's FSL (http://www.fmrib.ox.ac.uk/fsl/) using the "MNI Structural Atlas" and the "JHU White-Matter Tractography Atlas".

Database		FA	MD
A	1-NN	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	SVM	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	BP	0.75 (0.67-1.00)	0.78 (0.69-1.00)
	RBF	0.98 (0.97-1.00)	1.00 (1.00-1.00)
	PNN	1.00 (1.00-1.00)	0.54 (0.54-0.54)
	LVQ2	1.00 (1.00-1.00)	1.00 (1.00-1.00)
B	1-NN	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	SVM	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	BP	0.75 (0.66-1.00)	0.78 (0.70-1.00)
	RBF	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	PNN	1.00 (1.00-1.00)	0.52 (0.52-0.52)
	LVQ2	1.00 (1.00-1.00)	1.00 (1.00-1.00)
C	1-NN	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	SVM	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	BP	0.77 (0.68-1.00)	0.77 (0.68-1.00)
	RBF	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	PNN	1.00 (1.00-1.00)	0.52 (0.52-0.52)
	LVQ2	1.00 (1.00-1.00)	1.00 (1.00-1.00)
D	1-NN	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	SVM	1.00 (1.00-1.00)	1.00 (1.00-1.00)
	BP	0.77 (0.68-1.00)	0.77 (0.68-1.00)
	RBF	1.00 (1.00-1.00)	0.84 (0.79-0.90)
	PNN	0.99 (0.99-1.00)	0.55 (0.55-0.55)
	LVQ2	1.00 (1.00-1.00)	1.00 (1.00-1.00)
E	1-NN	0.94 (0.90-0.99)	1.00 (1.00-1.00)
	SVM	0.95 (0.90-1.00)	1.00 (1.00-1.00)
	BP	0.76 (0.67-1.00)	0.77 (0.68-1.00)
	RBF	0.92 (0.90-0.94)	0.89 (0.91-0.88)
	PNN	0.94 (0.90-0.99)	0.52 (0.52-0.52)
	LVQ2	0.97 (0.94-1.00)	1.00 (1.00-1.00)

Table II *10-fold cross-validation results. Accuracy (Sensitivity, Specificity)*

4. Results

The results are presented in table II. The most striking result is that we found optimal performance of almost all classifiers built from the provided feature vectors. The only exceptions were the results of PNN on MD data; tuning of the Gaussian kernel variance was more difficult than applying the training algorithm of other approaches . Also BP shows lower performance than the others. The second general result is that MD features seem to perform slightly better than FA features, disregarding the anomaly of PNN classifiers. In the experimental design we wanted to test if decreasing the size of the feature vectors had an impact on the classifiers performance. We found that performance was not affected down to the smallest feature vector (database E) where decreases in performance can be appreciated in all the classifiers for the FA data, while 1-NN, SVM and LVQ2 maintain their performance for MD data.

5. Conclusion

The goal of this paper was to test the hypothesis that classification algorithms constructed using statistical and Neural Network approaches can discriminate between schizophrenia patients and control subjects on the basis of features extracted from DTI data. The way to build the feature vectors has been the direct selection of voxels from the DTI-derived FA and MD scalar valued volumes that show a high correlation with the control variable that labels the subjects. The selected voxels roughly correspond to findings reported in the medical literature. Surprisingly, all the classifiers obtain near perfect results. Despite the simplicity of our feature extraction process, the results compare well with other results found in the literature [9, 34]. We think that appropriate pre-processing of the data is of paramount importance and can not be disregarded trusting that ensuing statistical or machine learning processes may cope with the errors introduced by lack of appropriate data normalization. Therefore, our main conclusion is that the proposed feature extraction is very effective in providing a good discrimination between schizophrenia patients that can easily be exploited by the classifier construction algorithms. The main limitation of this study is that the results come from a small database. Therefore, more extensive testing will be needed to confirm our conclusions. Nevertheless, we are making available [5] the actual data employed in the computational experiments to allow for independent validation of our results.

Acknowledgements

Thanks to the National Alliance for Medical Image Computing and the Brigham & Women's Hospital for making the database used for this study publicly available.

[5] http://www.ehu.es/ccwintco/index.php/GIC-experimental-databases

References

[1] N. Andreone, M. Tansella, R. Cerini, A. Versace, G. Rambaldelli, C. Perlini, N. Dusi, L. Pelizza, M. Balestrieri, C. Barbui, M. Nosè, A. Gasparini, and P. Brambilla. Cortical white-matter microstructure in schizophrenia. diffusion imaging study. *The British Journal of Psychiatry: The Journal of Mental Science*, 191:113–119, August 2007. PMID: 17666494.

[2] American Psychiatric Association. *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders (Diagnostic & Statistical Manual of Mental Disorders*. American Psychiatric Press Inc., 4th text revision edition, July 2000.

[3] B.B. Avants, C.L. Epstein, M. Grossman, and J.C. Gee. Symmetric diffeomorphic image registration with Cross-Correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, February 2008. PMID: 17659998 PMCID: 2276735.

[4] P.E. Barta, G.D. Pearlson, L.B. Brill, R. Royall, I.K. McGilchrist, A.E. Pulver, R.E. Powers, M.F. Casanova, A.Y. Tien, S. Frangou, and R.G. Petty. Planum temporale asymmetry reversal in schizophrenia: replication and relationship to gray matter abnormalities. *The American Journal of Psychiatry*, 154(5):661–667, May 1997. PMID: 9137122.

[5] P.J. Basser, J. Mattiello, and D. LeBihan. MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, 66(1):259–267, January 1994. PMID: 8130344 PMCID: 1275686.

[6] T.E.J. Behrens, M.W. Woolrich, M. Jenkinson, H. Johansen-Berg, R.G. Nunes, S. Clare, P.M. Matthews, J.M. Brady, and S.M. Smith. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine*, 50(5):1077–1088, 2003.

[7] A. Breier, R.W. Buchanan, A. Elkashef, R.C. Munson, B. Kirkpatrick, and F. Gellad. Brain morphology and schizophrenia: A magnetic resonance imaging study of limbic, prefrontal cortex, and caudate structures. *Archives of General Psychiatry*, 49(12):921–926, December 1992.

[8] D.S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

[9] A. Caprihan, G.D. Pearlson, and V.D. Calhoun. Application of principal component analysis to distinguish patients with schizophrenia from healthy controls based on fractional anisotropy measurements. *NeuroImage*, 42(2):675–682, August 2008.

[10] S. Chen, C.F.N. Cowan, and P.M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, 1991.

[11] Y. Fan, D. Shen, R.C. Gur, R.E. Gur, and Christos Davatzikos. COMPARE: classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging*, 26(1):93–105, January 2007. PMID: 17243588.

[12] M. Frederikse, A. Lu, E. Aylward, P. Barta, T. Sharma, and G. Pearlson. Sex differences in inferior parietal lobule volume in schizophrenia. *The American Journal of Psychiatry*, 157(3):422–427, March 2000.

[13] M. García-Sebastián, A. Savio, M. Graña, and J. Villanúa. On the use of morphometry based features for alzheimer's disease detection on MRI. In *Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence*, pages 957–964, Salamanca, Spain, 2009. Springer-Verlag.

[14] J.M. Goldstein, J.M. Goodman, L.J. Seidman, D.N. Kennedy, N. Makris, H. Lee, J. Tourville, V.S. Caviness, S.V. Faraone, and M.T. Tsuang. Cortical abnormalities in schizophrenia identified by structural magnetic resonance imaging. *Archives of General Psychiatry*, 56(6):537–547, June 1999.

[15] R.E. Gur, P.E. Cowell, A. Latshaw, B.I. Turetsky, R.I. Grossman, S.E. Arnold, W.B. Bilker, and R.C. Gur. Reduced dorsal and orbital prefrontal gray matter volumes in schizophrenia. *Archives of General Psychiatry*, 57(8):761–768, August 2000.

[16] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 2 edition, July 1998.

[17] D.P. Holinger, M.E. Shenton, C.G. Wible, R. Donnino, R. Kikinis, F.A. Jolesz, and R.W. McCarley. Superior temporal gyrus volume abnormalities and thought disorder in Left-Handed schizophrenic men. *The American Journal of Psychiatry*, 156(11):1730–1735, November 1999.

[18] T. Kohonen. Learning vector quantization. In *The handbook of brain theory and neural networks*, pages 537–540. MIT Press, 1998.

[19] M. Kubicki, R. McCarley, C.-F. Westin, H-J Park, S. Maier, R. Kikinis, F.A. Jolesz, and M.E. Shenton. A review of diffusion tensor imaging studies in schizophrenia. *Journal of psychiatric research*, 41(1-2):15–30, 2007. PMID: 16023676 PMCID: 2768134.

[20] M. Kubicki, H. Park, C.-F. Westin, P.G. Nestor, R.V. Mulkern, S.E. Maier, M. Niznikiewicz, E.E. Connor, J.J. Levitt, M. Frumin, R. Kikinis, F.A. Jolesz, R.W. McCarley, and M.E. Shenton. DTI and MTR abnormalities in schizophrenia: Analysis of white matter integrity. *NeuroImage*, 26(4):1109–1118, July 2005. PMID: 15878290 PMCID: 2768051.

[21] M. Kubicki, C.F. Westin, R.W. McCarley, and M.E. Shenton. The application of DTI to investigate white matter abnormalities in schizophrenia. *Annals of the New York Academy of Sciences*, 1064(1):134–148, 2005.

[22] M. Kyriakopoulos, T. Bargiotas, G.J. Barker, and S. Frangou. Diffusion tensor imaging in schizophrenia. *European Psychiatry*, 23(4):255–273, June 2008.

[23] P.K. McGuire and C.D. Frith. Disordered functional connectivity in schizophrenia. *Psychological Medicine*, 26(4):663–667, July 1996. PMID: 8817700.

[24] M. Niznikiewicz, R. Donnino, R.W. McCarley, P.G. Nestor, D.V. Iosifescu, B.O'Donnell, J. Levitt, and M.E. Shenton. Abnormal angular gyrus asymmetry in schizophrenia. *The American Journal of Psychiatry*, 157(3):428–437, March 2000.

[25] C. Pierpaoli, P. Jezzard, P.J. Basser, A. Barnett, and G. Di Chiro. Diffusion tensor MR imaging of the human brain. *Radiology*, 201(3):637–648, December 1996. PMID: 8939209.

[26] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. *Learning internal representations by error propagation*, pages 318–362. MIT Press, 1986.

[27] T.E. Schlaepfer, G.J. Harris, A.Y Tien, L.W. Peng, S. Lee, E.B. Federman, G.A. Chase, P.E. Barta, and G.D. Pearlson. Decreased regional cortical gray matter volume in schizophrenia. *The American Journal of Psychiatry*, 151(6):842–848, June 1994.

[28] M.E. Shenton, C.C. Dickey, M. Frumin, and R.W. McCarley. A review of MRI findings in schizophrenia. *Schizophrenia research*, 49(1-2):1–52, April 2001. PMID: 11343862 PMCID: 2812015.

[29] M.E. Shenton, R.Kikinis, F.A. Jolesz, S.D. Pollak, M. LeMay, C.G. Wible, H. Hokama, J. Martin, D. Metcalf, M. Coleman, and R.W. McCarley. Abnormalities of the left temporal lobe and thought disorder in schizophrenia. *New England Journal of Medicine*, 327(9):604–612, 1992.

[30] S.M. Smith, M. Jenkinson, M.W. Woolrich, C.F. Beckmann, T.E.J. Behrens, H.Johansen-Berg, P.R. Bannister, M. De Luca, I. Drobnjak, D.E. Flitney, R.K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J.M. Brady, and P.M. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(Supplement 1):S208–S219, 2004.

[31] D.F. Specht. Probabilistic neural networks. *Neural Networks*, 3(1):109–118, 1990.

[32] P.R. Szeszko, R.M. Bilder, T. Lencz, S. Pollack, J.M. Alvir, M. Ashtari, H. Wu, and J.A. Lieberman. Investigation of frontal lobe subregions in first-episode schizophrenia. *Psychiatry Research*, 90(1):1–15, February 1999. PMID: 10320207.

[33] V.N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, September 1998.

[34] P. Wang and R. Verma. On classifying Disease-Induced patterns in the brain using diffusion tensor images. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2008*, pages 908–916. 2008.

[35] C. Wernicke. *Grundriss der Psychiatrie in klinischen Vorlesungen / von Carl Wernicke*. VDM Verlag Dr. Müller, Saarbrücken, 2007.

[36] C.G. Wible, M.E. Shenton, H. Hokama, R. Kikinis, F.A. Jolesz, D. Metcalf, and R.W. McCarley. Prefrontal cortex and schizophrenia: A quantitative magnetic resonance imaging study. *Archives of General Psychiatry*, 52(4):279–288, April 1995.

[37] U. Yoon, J.-M. Lee, K. Im, Y.-W. Shin, B.H. Cho, I.Y. Kim, J.S. Kwon, and S.I. Kim. Pattern classification using principal components of cortical thickness and its discriminative pattern in schizophrenia. *NeuroImage*, 34(4):1405–1415, February 2007.