Using fMRI to Diagnose Schizophrenia

Department of Computing Science
University of Alberta

Department of Computing Science
University of Alberta

Shashindra Silva
Department of Electrical and Computer Engineering
University of Alberta
jayamuni@ualberta.ca

Abstract

Diagnosis of schizophrenia is a challenging task for which diagnostic tests have yet to be developed [6]. Although Functional Magnetic Resonance Imaging (fMRI) methods have become more common in the diagnosis of mental disorders have become more popular, for schizophrenia diagnosis fMRI methods need to be more robust and reliable. Similar studies [9][10] have shown that fMRI can be used in conjunction with Sparse Gaussian Markov Random Field (SGMRF) to produce high accuracy in diagnosis of illness. However having a dataset with homogeneous distribution of illness makes this result less reliable and creates the need for more evidence using heterogeneous dataset in terms of illness. In this work we pursue two paths to tackle this problem. First, we evaluate performance of Sparse Gaussian Markov Random Field (SGMRF) on fMRI data brain scans, and second we study on Regions of Interest (ROI) as defined by Power *et al.* [8]. We used 5 fold cross validation for hyper parameter tuning and 20% hold-out set for test. Accuracies that We have obtained the following accuracies using this method: for whole brain features and - for ROI features. While these result are slightly less than the results obtained by Rish *et al.*, they are on par with Rosa *et al.* results.

1 Introduction

Schizophrenia is a mental/psychiatric disorder [9, 5] known affect blood flow in the brain [5] where those who are affected can experience hallucinations, delusions and diminished mental capacities to varying extents [4]. While several features of schizophrenia have proven useful for its diagnosis there are current no set of features that have sufficient sensitivity or specificity to be used in diagnostic tests [4]. This effectively means that subjectivity plays a role when a physician is diagnosing a patient.

Functional Magnetic Resonance Imaging (fMRI) is a tool for recording functional changes caused by neuron activity[]. When a person is doing a task, neuron activity fluctuates and in order to provide the energy needed for this activity, the blood flow increases to supply the neurons with the needed glucose, which is not stored in the brain[]. More blood flow also brings more oxygen through blood vessels. This change in the level of oxygenated blood known as oxyhemoglobin and deoxyhemoglobin (oxygenated or deoxygenated blood) changes the magnetic susceptibility of blood (BOLD signal) which is detectable through fMRI [].

fMRI is one of the most used and efficient tools in the study of psychiatric disorders such as Schizophrenia[]. An advantage of fMRI in medical diagnosis is that it is non-invasive. This means

that unlike some other imaging methods, no instruments or dyes are placed in the patients body body, this method operates without using them[].

One of the approaches that has been used for studying schizophrenia is Sparse Gaussian Markov Random Field(SGMRF) [9][10]. The primary advantage of using this method is that the functional network of the brain can be captured using the precision matrix [9]. By using the resulting network, healthy subjects can be differentiated from schizophrenic ones, by observing differences in the functional connectivity of the brain. Currently automated approaches to schizophrenia diagnosis have been able to yeild accuracies of 93% for data that originates from a single location [9] and up to approximately 80% for data that originates from multiple locations [2].

In this work we consider

The rest of the paper is organized as follows.

2 Background and Prior Work

2.1 Regions of Interest and Single-Voxel Analysis

There are two main approaches for extracting information from fMRI images. The first is a single-voxel approach and the second is to study regions of interest (ROI) [3]. The trade-off between these two approaches is that a single-voxel approach requires the analysis of every voxel and is subject to the low signal-to-noise ratios of individual voxels, whereas a region based approach is only effective if the selected regions capture all relevant information in an fMRI task [3]. In 2011, Power *et al.* identified 264 putative function regions of interest derived from resting state fMRI, where no specific task being performed during data collection [8]. JD Power argues in his video abstract that these regions are currently the best representation of functional networks in the brain that are available [8].

2.1.1 Calculating Degrees

When analyzing fMRI data, features such as voxel degrees can be extracted for use with a machine learner. Voxel degrees represent the connectedness of voxels in the brain with the other voxels and are described as "the number of voxel neighbours in a network" [9]. Degrees are calculated by performing multiple Pearson correlation comparisons between the i^{th} voxel and every other voxel. Once correlation values have been determined, a threshold is applied to the correlation matrix. This results in binary matrix where 1 represents a correlation value above the threshold and 0 represents a value below. Finally, for each voxel the number of 1 entries are summed (excluding the comparison against itself) and this becomes the degree of the voxel.

2.2 Multi-site Comparisons

Although, there were several prior work on the classification of schizophrenia patients using fMRI, most of them are based on data from a single site. Classification from multi-site data is inherently difficult due to the batch effects resulted from the use of different machines and environments. At the same time, multi-site analysis can easily be generalized for a new data set from a totally different source. Cheng et al. [2] analyzes a multi site data set which is obtained from five different sites with different machines. They used SVM for classify schizophrenia patients and healthy controls and obtained accuracies in the range of $73.53 - 80.92\%$. In this research we will try to obtain results with similar accuracies, but using probabilistic graphical methods.

2.3 Principal Component Analysis

2.4 Support Vector Machines

In a 1995 paper by Vladimir Vapnik *et al.* the concept of support Vector Machines (SVMs) was introduced as a statistical tool for classification problems [11]. In our work we use the most simple SVM, a linear SVM, because it does not use non-linear kernels and therefore has no hyperparameters that need to be tuned with cross validation. Instead, a linear SVM learns the "maximum-margin hyperplane" classifier which is a linear combination of the input features and partitions the data space into separate classes. The term "maximum-margin" refers to the SVM's ability to find the maximal

separation between classes and the hyperplane, therefore creating the largest "margin" [11]. We can be guaranteed of the optimality of the result as the SVM problem is known to be convex [1]. The simplest form of a linear SVM minimizes $\frac{1}{2}||w||^2$ subject to the constraint $y_i(x_i w + b) - 1 \geq 0, \forall i$ where w is the weight vector, x_i is the instance's features, y_i is the instance label and b is the bias term of the model. Unfortunately, this version of the SVM only works in the case where the data is linearly separable. To extend the SVM to linearly inseparable data, the addition of positive slack variables is required, such that the constraints become $\forall i \; y_i(x_i w + b) - 1 + \xi_i \geq 0$ and $\xi_i \geq 0$, where xi_i is the slack variable for the instance i [1].

2.5 SGMRF

One variation of Markov Random Field is Gaussian Random which is generally used for continuous space of variables and has well-defined mathematic properties that can be computed. Multivariate Gaussian density function over a set of random variables X is defined as below:

$$p(X) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(X-\mu)^t \Sigma^{-1}(X-\mu)\right\}. \tag{1}$$

In the equation μ is the mean and Σ is the covariance matrix. We can set μ to zero and replace Σ^{-1} with C the equation (1) which can be written as the following form. It also should be noted that $C = \Sigma^{-1}$ and is known as the precision matrix in the literature.

$$p(X) = (2\pi)^{-n/2}|C|^{1/2} \exp\left\{-\frac{1}{2}X^t C X\right\}. \tag{2}$$

It has been shown by [**?**] that X_i and X_j are independent if and only if their corresponding entries in the precision matrix (C) are zero. It can be concluded that missing edge in the MRF will lead to zero entries in the precision matrix[**?**]. The Problem of learning probabilistic graphical model for a given dataset is reduced to learning the precision matrix given this proof.

The Log-likelihood of the dataset, assuming each row of the data is a p-dimensional vector and consisting n samples $D = \{X_1, X_2, ..., X_n\}$, can be written as follow. It should also be noted that we assume each sample is identically independent distributed(iid).

$$L(D) = \frac{n}{2}|C| - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^T C (X_i - \mu) + const, \tag{3}$$

Here, const is a constant which is not dependant on μ or C. We also can center the data in such a way that $\mu = 0$, so the second term reduces to $\frac{1}{2}\sum_{i=1}^{n} X_i^T C X_i$ which is equal to $\frac{n}{2}$ tr(AC), where tr is trace of the matrix. The above formula will can be written as shown in equation (4) where the log-likelihood maximization problem is shown in equation (5).

$$L(D) = \frac{n}{2}[|C| - |tr(AC)|] + const \tag{4}$$

$$\max_{C \succ 0} |C| - tr(AC) \tag{5}$$

Where $|...|$ represents determinant and A is the empirical covariance matrix calculated by $A = \frac{n}{2}\sum_{i=1}^{n} X_i^T X_i$, or maximum likelihood estimation fo Σ. It should also be noted the $C \succ 0$ constraint makes C positive definite.

sparsity formulation for Gaussian Markov Random Field

If we impose the Laplace prior on the precision matrix $p(C_{ij}) = \frac{\lambda_{ij}}{2} e^{-\lambda_{ij}|C_{ij}|}$ we can achieve sparsity. This will change the equations (4) and (5) to the following form:

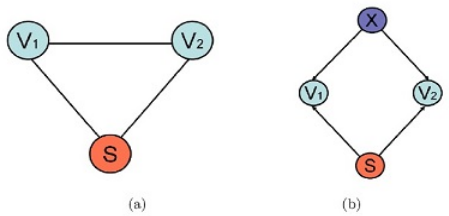

Figure 1: Graphical models of voxel interactions Rish *et al.*

$$L(D) = \frac{n}{2}[ln|C| - |tr(AC)|] - \lambda||C||_1, \quad (6) \qquad \max_{C \succ 0} ln|C| - tr(AC) - \rho||C||_1, \quad (7)$$

where l_1-norm of C is: $||C||_1 = \sum_{ij} C_{ij}$, and $\rho = \frac{n}{2}\lambda$.

One obvious way for learning the parameters for joint distribution probability is using regularized likelihood maximization such as AIC and BIC. However, finding the simplest model which fits the data is a NP-hard problem using this approach. There are also few limitations that make using this approach less desirable. First Empirical covariance matrix may not even exist, specially when the number of features in the dataset is more than the number of samples. Which is the case in fMRI studies. fMRI datasets usually have fraction of samples to features. Second problem using this approach is not having zero elements in the inverse of empirical covariance matrix. Hence, to construct the MRF using this matrix one should include explicit sparsity constraint.[**?**]

These two major problems cause us to search for other solution to build the precision matrix given the dataset. Fortunately one can use the alternative approximation approach for the above problem. Recently there have been some successes for getting the approximate precision matrix given the dataset. Glasso[**?**], block-coordinate descend(BCD) known as COVSEL and projected gradient approach are among them. For further reading on these methods readers can refer to [**?**]. In this study we are using Varsel and Glasso as the preferred method for obtaining the precision matrix and subsequently MRF.

2.6 Rish *et al.*

In this work several approaches have been experimented. The most successful ones are SVM and MRF classifiers. For the MRF classifier they have used the same methods that has been mentioned in the *SGMRF* section. The error rate for SVM was 7% using pairwise correlation features, and 14% for the MRF using degree 100 most significant voxels in the degree feature set. Here we focus on their approach for MRF classifier.

In the MRF classifier, first they split training data in two different group. For each group they compute the precision matrix(assuming the data is centered unless μ needs to be computed as well).

3 Methodology

3.1 Data Set

Data used for this study were downloaded from the Function BIRN Data Repository (http://fbirnbdr.birncommunity.org:8080/BDR/). The original data contained nine sites and 235 subjects. However, during the preprocessing steps some of the subjects were removed because they could not be preprocessed properly, resulting in a final set of 95 subjects and five sites. Data were preprocessed by Dr. Mina Gheiratmand by using FSL software

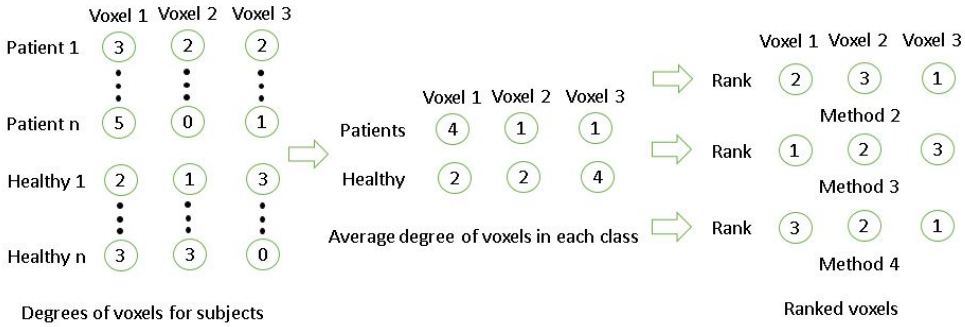

Figure 2: Voxel ranking methods to select best voxels. In step 2, the differences of the average degrees between patients and healthy subjects are 2, -1, and -3 respectively for the three voxels.

(http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). Our data contained 46 schizophrenic subjects and 49 healthy subjects. Each subject had four runs (repeated scans) so effectively our data set had 380 subjects. Each run had 137 time slices and each time slice had the signal amplitude of over $100,000$ voxels. The voxels were referred using the 3D coordinates and thus the dimensionality of a single run was $\mathcal{R}^{N_1 \times N_2 \times N_3 7 \times 137}$, wh We removed 80% of subjects from each site as our holdout set and used the remaining set as our training set. Furthermore, we made sure that the holdout set was also balanced, so that the ratio between the patients and healthy subjects was same in the holdout and training sets for each site. We made sure that all the runs from the same subject either belonged to the holdout set or to the training set. The training set was used for finding the hyperparameters of the system using five-hold cross validations. After finding the hyper parameters we used the full training set to train the system with the selected hyper parameter and obtained the accuracy of the hold out set. We repeated this procedure five times for different hold out sets to get an average accuracy.

3.2 SGMRF with log degrees of ranked voxels

First, we executed the approach proposed by Rish *et al.* [9] on our data set. We had degrees of each voxels as well as the smoothed log values of the voxels for each subject. Furthermore, as some voxels of some subjects had zero values for the BOLD signal, an universal mask was used to select the voxels which had a non-zero BOLD signal value through the whole data set. This resulted in a 28719 voxels per each subject. However, using all the available voxels resulted in significantly increased computation time and thus selecting the best voxels (based on the training data) to separate the healthy and schizophrenic subjects was very important.

To select the best voxels out of the possible 28719 voxels, we experimented with multiple approaches. These included:

1. Selecting voxels based on t-tests.
2. Selecting voxels based on absolute differences of mean degrees of a voxel between schizophrenic and healthy subjects.
3. Selecting voxels based on differences of mean degrees of a voxel between schizophrenic and healthy subjects.
4. Selecting voxels based on differences of mean degrees of a voxel between healthy and schizophrenic subjects.

The last three voxel ranking methods are explained in Figure 3.2. Out of the above four approaches we obtained the highest accuracy on our cross validation set by the third approach. Also the number of voxels k to be used in the SGMRF model was decided by running the experiment for different k values and finding that $k = 20$ performed well during cross validation. By using the selected voxels, we obtained the precision matrices for schizophrenic and healthy subjects. Similar to the value of k, the sparsity coefficient λ was also obtained through a hyper parameter search on our cross validation set.

3.3 SGMRF with degrees from Dorsolateral Prefrontal Cortex area

Connections in certain areas of the brain specifically dorsolateral prefrontal cortex (DLPFC) [7] and thalamocortical [2] circuitry. Thus, we extracted the degrees of voxels in these areas and used them to build the SGMRF model. Since we had only 4504 voxels for each subject, we did not perform any voxel rankings in this approach. Similar to the previous approach using the smoothed log values of the degrees resulted in higher accuracy.

3.4 Methods using Power *et al.*'s ROIs

The following three methods used in this subsection all involve the 264 regions of interest that are described in work by Power *et al.*. Due to missing data resulting from incomplete scans of subjects, 8 regions were removed from analysis leaving 256 ROIs. The regions removed are characterized by region number with the associated MNI coordinates described in Power *et al.* provided in parenthesis: 81 (-44, 12, -34), 82 (46, 16, -30), 128 (52, 7, -30), 184 (17, -80, -34), 247 (33, 12, -34), 248 (-31, -10, -36), 249 (49, -3, -38) and 250 (-50, -7, -39). We use the approach described by Vega *et al.* [12] to summarize regions of interest by taking region averages for each of the $5mm$ radius spheres that define regions. Finally this resulted in a 137×256 matrix for each subject where each row is a time point across all regions and each column is the average time series for a single region.

3.4.1 ROI with Subject Concatenation

For this approach we follow the method described by Vega *et al.*. Training set subjects in each class are first concatenated so that two large matrices are created, one with dimension $ns * 137 \times 256$ and the other with dimension $nh * 137 \times 256$ where ns is the number of schizophenic subjects in the training set and nh is the number of health subjects. To make learning the SGMRF structure easier, feature normalization was performed where for each class and each region in that class, the region mean was subtracted from each time point and each time point was divided by the region standard deviation.

$$ClassRegion_i = \frac{ClassRegion_i - mean(ClassRegion_i)}{std(ClassRegion_i)} \tag{8}$$

Next, with the use of Glasso, a SGMRF structure was learned for each class which resulted in two sparse precision matrices that encode the independences learned.

Finally, when presented with a new subject from the hold-out set, the equation below was used to determine the likelihood of the subject belonging to each class. The class with the highest likelihood then became the predicted label for the subject.

A modified version of this approach was also implemented and tested but has been omitted for brevity and due to it obtaining poor results. In this variant, a Fourier transform was used on each of the subject's time series data to obtain Fourier coefficients. These Fourier coefficients were then used in place of the original time series for learning a classifier.

$$inserthere \tag{9}$$

3.4.2 Region Degrees and SVMs

Like the work described in Rish *et al.*, region degrees are also used in this approach but in this case, we only considered the degrees that result from the 256 ROIs. This process is illustrated in Figure 3. For each subject we generated a correlation matrix containing the pair-wise correlations between all of average time series for each region. Because we were not interested in the correlation between a region and itself (which is trivially 1) we subtracted the diagonal of the correlation matrix from itself. Using a threshold of 0.7, which was chosen based on work by Rish *et al.*, we then proceeded to create the binary matrix and sum the column values as described previously.

Now that we had a vector of region degrees (1×256) for each subject, a SGMRF was used to again to build a classifier. This followed similarly from the previous approach except that now the input to Glasso was a $ns \times 256$ matrix for schizophrenic patients and a $nh \times 256$ matrix for healthy

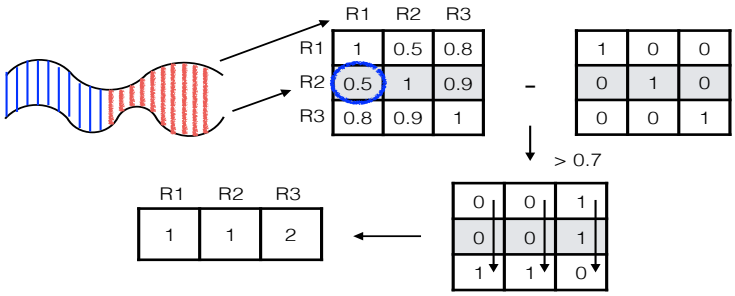

Figure 3: Pair-wise comparisons are made between region time series data to produce a correlation matrix. The diagonal is removed and a threshold is applied to binarize the matrix. Sums are collected for each region to produce the "degree" of that region.

subjects. Again this resulted in two 256×256 precision matrices which we could use in likelihood calculations for subjects from the hold-out set.

Additionally, we also used a linear SVM classifier trained directly on region degree data to compare its performance to the SGMRF classifier.

3.4.3 Individual MRF Structure Classification

This approach is similar to the region concatenation approach described previously except that here we did not concatenate subjects and instead learned a precision matrix (SGMRF structure) for each subject individually. For example, if we had ns_{total} schizophrenics in our dataset and nh_{total} healthy subjects ($ns_{total} + nh_{total}$ 137×256 matrices) then we would use Glasso to generate $ns + nh$ precision matrices.

To build a classifier, a linear SVM is trained on the precision matrices generated from the training set and then tested on the precision matrices generated from the hold-out set.

3.5 Your Approach? Farhad

Describe your experiments

4 Results and Discussion

	Rish (Full)	Rish (DLPFC)	ROI + Concatenation	Region Degrees	MRF structure	PCA
Classifier	SGMRF	SGMRF	SGMRF	SVM	SVM	1
Accuracy	72.32%	63.52%	74.17	60.65	69.44	1

Table 1: Accuracy on 5 Striated Hold-out Sets

As Seen in Table 1 methods using a SGMRF classifier were able to obtain the best results, of which 74.17 was the highest average accuracy recorded. This seems to suggest that a SGMRF seems to be a more effective tool for this particular problem. Because the SGMRF is a generative tool whereas the SVM is discriminative, it may be the case that the SVM is simply unable to capture the differences in network structure between patients and controls and it is these network differences that define schizophrenia as a disease. When comparing the ROI approch with patient concatenation and the Rish approach using all voxels against the Rish approach that only considers the DLPFC regions we see the former two get much better results. This seems to indicate that the DLPFC region is not sufficient to capture the difference between patients and controls. Another noteworthy difference is that of the differences between the approach creating degree features from the region averages and the other better performing methods. This discrepancy may be due

to (Not sure

5 Conclusions and Future Work

In this work we studied the use of fMRI and machine learning to diagnose schizophrenia, given a set of healthy and schizophrenic subjects. We found that a region based approach using SGMRFs yielded the highest accuracy (74.16%) when averaged over 5 hold-out sets where the hold-out set was composed of 20% of each site used in the study.

6 Acknowledgements

We would like to thank Dr. Mina Gheiratmand for preprocessing the fMRI data and co-coaching our project, Sugai Liang for her advice and Roberto Vega for all his advice and help. Finally, we would also like to thank Dr. Irina Rish for the feature extraction code she provided and Dr. Greiner for his supervision of this project.

References

[1] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[2] Wei Cheng, Lena Palaniyappan, Mingli Li, Keith M Kendrick, Jie Zhang, Qiang Luo, Zening Liu, Rongjun Yu, Wei Deng, Qiang Wang, Xiaohong Ma, Wanjun Guo, Susan Francis, Peter Liddle, Andrew R Mayer, Gunter Schumann, Tao Li, and Jianfeng Feng. Voxel-based, brain-wide association study of aberrant functional connectivity in schizophrenia implicates thalamocortical circuitry. *Npj Schizophrenia*, 1, May 2015.

[3] Ruth Heller, Damian Stanley, Daniel Yekutieli, Nava Rubin, and Yoav Benjamini. Cluster-based analysis of fmri data. *NeuroImage*, 33(2):599–608, 2006.

[4] Assen Jablensky. The diagnostic concept of schizophrenia: its history, evolution, and future prospects. *Dialogues Clin Neurosci*, 12(3):271–287, 2010.

[5] Nomura Kenji. Pressure-induced performance decrement in verbal fluency task through prefrontal overactivation: A near-infrared spectroscopy study. *Front. Neurosci.*, 4, 2010.

[6] Philip McGuire, Oliver D Howes, James Stone, and Paolo Fusar-Poli. Functional neuroimaging in schizophrenia: diagnosis and drug discovery. *Trends in Pharmacological Sciences*, 29(2):91 – 98, 2008.

[7] S G Potkin, J A Turner, G G Brown, G McCarthy, D N Greve, G H Glover, D S Manoach, A Belger, M Diaz, C G Wible, J M Ford, D H Mathalon, R Gollub, J Lauriello, D O'Leary, T G M van Erp, A W Toga, A Preda, K O Lim, and FBIRN. Working memory and DLPFC inefficiency in schizophrenia: The FBIRN study. *Schizophrenia Bulletin*, 35(1):19–31, November 2008.

[8] Jonathan D. Power, Alexander L. Cohen, Steven M. Nelson, Gagan S. Wig, Kelly Anne Barnes, Jessica A. Church, Alecia C. Vogel, Timothy O. Laumann, Fran M. Miezin, Bradley L. Schlaggar, and Steven E. Petersen. Functional network organization of the human brain. *Neuron*, 72(4):665–678, nov 2011.

[9] Irina Rish, Guillermo Cecchi, Benjamin Thyreau, Bertrand Thirion, Marion Plaze, Marie Laure Paillere-Martinot, Catherine Martelli, Jean-Luc Martinot, and Jean-Baptiste Poline. Schizophrenia as a network disease: Disruption of emergent brain function in patients with auditory hallucinations. *PLoS ONE*, 8(1):e50625, jan 2013.

[10] Maria J. Rosa, Liana Portugal, John Shawe-Taylor, and Janaina Mourao-Miranda. Sparse network-based models for patient classification using fMRI. In *2013 International Workshop on Pattern Recognition in Neuroimaging*. Institute of Electrical & Electronics Engineers (IEEE), jun 2013.

[11] Armin Shmilovici. Support vector machines. In *Data Mining and Knowledge Discovery Handbook*, pages 257–276. Springer, 2005.

[12] Roberto Vega, Khare Kriti, and Sayem Mohammad Siam. Gender and age group classification using functional magnetic resonance imaging and gaussian markov random fields. *Unpublished*, 2015.