A Non-parametric Approach to Detect Epileptogenic Lesions Using Restricted Boltzmann Machines

Yijun Zhao
CS Department
Tufts University
yzhao@cs.tufts.edu

Bilal Ahmed
CS Department
Tufts University
bahmed01@cs.tufts.edu

Thomas Thesen
Comprehensive Epilepsy Center
New York University
thomas.thesen@med.nyu.edu

Karen E. Blackmon
Comprehensive Epilepsy Center
New York University
karen.blackmon@nyumc.org

Jennifer G. Dy
ECE Department
Northeastern University
jdy@ece.neu.edu

Carla E. Brodley
College of Computer and Information Science
Northeastern University
c.brodley@neu.edu

ABSTRACT

Visual detection of lesional areas on a cortical surface is critical in rendering a successful surgical operation for Treatment Resistant Epilepsy (TRE) patients. Unfortunately, 45% of Focal Cortical Dysplasia (FCD, the most common kind of TRE) patients have no visual abnormalities in their brains' 3D-MRI images. We collaborate with doctors from NYU Langone's Comprehensive Epilepsy Center and apply machine learning methodologies to identify the resective zones for these *MRI-negative* FCD patients. Our task is particularly challenging because MRI images can only provide a limited number of features. Furthermore, data from different patients often exhibit inter-patient variabilities due to age, gender, left/right handedness, etc. In this paper, we introduce a new approach which combines the restricted Boltzmann machines and a Bayesian non-parametric mixture model to address these issues. We demonstrate the efficacy of our model by applying it to a retrospective dataset of MRI-negative FCD patients who are seizure free after surgery.

Keywords

mixture models, Bayesian non-parametric, Restricted Boltzmann Machine, predictive medicine, semi-supervised learning and application

1. INTRODUCTION

Epilepsy is a common neurological disorder, affecting approximately 1% of the population [14]. It is characterized by profound abnormal neural activity during seizures and interictal (between seizures) periods. Uncontrolled epilepsy can have harmful effects on the brain and has increased risk of injuries and sudden death [3]. About one third of epilepsy patients remain resistant to medical treatment [21]. Our research addresses the identification of lesions in the MRI's of patients with focal cortical dysplasia (FCD), which is recognized as the most common source of pediatric epilepsy [3, 35] and the third most common source in adults having medically intractable seizures [20, 22]. Early detection and subsequent surgical removal of the FCD lesion area is the most effective and is often the last hope for these patients.

The most widely used technology in identifying the epileptic lesions is MRI coupled with intracranial EEG (iEEG). For *MRI-positive* patients, (i.e., patients with visible abnormal areas in the MRI), the placement of electrodes on the cortex is informed by the pinpointed problematic regions detected by visual inspection of the MRI. However, for *MRI-negative* patients there is no visible lesion to guide precise electrode implantation [17]. The final target for surgical resection is based on both of these findings if available coupled with other clinical data. Consequently, the post-surgical success (i.e., seizure-freedom after surgery) ratio of MRI-positive to MRI-negative patients is 66% to 29%. Unfortunately, 45% of FCD patients are MRI-negative [37]. For this reason the surgical resection procedure remains highly underutilized as most practitioners are unwilling to operate in the absence of a visually detected lesion [34].

The machine learning task is to detect the lesional region(s) in MRI-negative patients. Specifically, our model identifies the abnormal areas in a patient's brain which in turn serve as a focus of attention mechanism for the neuroradiologists in placing the iEEG sensors on the patient's cortex. The work presented in this paper adopts a new Bayesian non-parametric approach as compared to our previous logistic regression (LR) based model [1]. Our LR model has been in use at NYU's Comprehensive Epilepsy Center since 2013. The new approach, as we present in Section 6, achieves improved performance compared to the LR model.

Our training data comes from the 3D-MRI images of MRI-negative FCD patients who underwent resective brain surgery at NYU and were seizure free after surgery. Furthermore, the resected tissue was histopathologically verified to contain FCD. We also have access to the MRI's of healthy controls who underwent the same MRI protocol. One challenge in applying machine learning to this dataset is the paucity

KDD '16, August 13-17, 2016, San Francisco, CA, USA

 ISBN 978-1-4503-4232-2/16/08…$15.00

DOI: http://dx.doi.org/10.1145/2939672.2939705

of features available from the MRI that are predictive of lesional tissue (see Section 2.1 for a complete list of the available features). To address this issue, we employ the Restricted Boltzmann Machines (RBMs) [32, 40] to learn a new set of nonlinear features. An RBM is an undirected graphical model [18] which can be interpreted as a stochastic neural network. Through training, an RBM learns a closed-form distribution of the input data. The units in the bottom layer of the network correspond to the observable attributes and the top layer consists of hidden units that can be viewed as nonlinear feature detectors [16]. Thus, an RBM is often employed to extract more meaningful features from input data [29].

Another challenge while learning with this dataset is the impact of inter-patient variability, i.e., each human brain has its own characteristics depending on the severity of the disease, age, gender, left/right handedness, lifestyle and genetics many of which are not uniformly available in our data. Thus, learning from data collected from all patients does not lead to satisfactory performance. To address this issue, we partition the data into subgroups with the goal that each subgroup will contain instances with similar brain characteristics. Because we do not have sufficient domain knowledge to estimate the number of subgroups, we employ a Bayesian non-parametric Dirichlet process mixture model (DPM, [33]) to infer the number of clusters automatically from the data.

Our proposed approach leads to an infinite mixture of RBMs. Because a typical mixture (finite or infinite) of RBMs is computationally intractable [23], we propose a two-stage method. We first apply an RBM to the entire dataset and cluster the patients based on the hidden layer of this model using a DPM. The second stage applies a weighted version of RBM to each non-empty cluster using the weights obtained from the DPM model. We discuss the rationale behind this two-stage method in Section 5.1.

We evaluate our model on MRI-negative patients from NYU who were seizure free after resective surgery. Our evaluation includes a successful detection indicator (Y or N) within the resected region and the performance (true negative rate, TNR) outside the resection zone. We need to measure the performance of the two regions separately because we not only want to locate the lesion but also don't want to misclassify non-lesional benign tissue. Note that, instead of using the true positive rate (TPR), we use an indication to measure whether there is any lesion detected inside the resection zone. This is because in the absence of any visual information to locate the lesion precisely, very generous margins are employed during the surgery to ensure that the patient is seizure free afterwards. As a result, the TPR is not a meaningful evaluation metric because the resection zone is much larger than the actual lesion and contains benign areas. Instead, we are interested in assessing whether there is any detection that can potentially guide the placement of iEEG sensors with the ultimate goal of reducing the size of the resection zone, which will lead to a much safer yet effective surgery. In our experiments, our model has accuracies of greater than 99.4% in the benign region and a successful detection in the *resection region* in 7 out of 12 (i.e., 58%) patients (see details in Section 6). This is in contrast to neuroradiologists' visual MRI inspections that have a detection rate of 0 out 12.

In this paper, we first describe our method of constructing training data from 3D-MRI images of human brains using surface morphometry [10, 12]. We then present a brief introduction to RBM and DPM models in Sections 3 and 4 respectively. In Section 5, we outline our model which integrates the RBM and DPM algorithms. We present and discuss our results in Section 6 and conclude in Section 7.

2. SURFACE BASED MORPHOMETRY

Surface based morphometry (SBM) provides the means to characterize and analyze the human brain by explicitly modeling the cortex using a suitable geometric model [10]. The cortical surface represents the outer layer of the brain modeled as a folded two-dimensional surface. It is extracted by delineating the boundary between the gray and white matter using T1-weighted MRI images [10]. The reconstructed surface is represented as a triangulated surface [10], and at each vertex on the surface different morphological features such as cortical thickness, curvature, sulcal depth, etc., can be calculated to characterize the cortex. Similarly, different morphological transforms can be applied to register the cortical surface to a standard surface also known as a group-atlas. Registration is achieved by aligning specific sulcal and gyral landmarks across the *reconstructed* cortical surfaces allowing for a precise comparison of individual cortical structures across subjects [11]. SBM has been used successfully for analyzing and detecting neurological abnormalities in various neurological disorders such as schizophrenia [28], autism [25], and epilepsy [35, 17].

2.1 Feature Extraction

In this work we use five features to represent each vertex on the reconstructed surface:

1. *Cortical thickness* represents the thickness of the cortex which is defined as the distance between the gray/white matter boundary and the outermost surface of the gray matter (pial surface). It is calculated at each vertex using an average of two measurements [11]: (a) the shortest distance from the white matter surface to the pial surface; and (b) the shortest distance from the pial surface at each point to the white matter surface.

2. *Gray/white-matter contrast (GWC)* represents the degree of blurring at the gray/white-matter boundary. GWC is estimated by calculating the non-normalized T1 image intensity contrast at 0.5mm above and below the gray/white boundary with trilinear interpolation of the images. The range of GWC values lies in $[-1, 0]$, with values near zero indicating a higher degree of blurring of the gray/white boundary.

3. *Sulcal depth* characterizes the folded structure of the cortex. It is estimated by calculating the dot product of the movement vectors with the surface normal [12], and results in the calculation of the depth/height of each point above the average surface. The values of sulcal depth lie in the range $[-2, 2]$ with lower values indicating a location in the sulcus whereas higher values indicate a location on the gyral crown.

4. *Curvature* is measured as $\frac{1}{r}$, where r is the radius of an inscribed circle and mean curvature represents the average of two principal curvatures with a unit of 1/mm

[27]. Mean curvature quantifies the sharpness of cortical folding at the gyral crown or within the sulcus, and can be used to assess the folding of small secondary and tertiary folds in the cortical surface.

5. *Jacobian distortion* measures the magnitude of the nonlinear transform needed to wrap each vertex on the subject's brain to a target vetex on the average surface, as part of the registration process. This nonlinear transform is needed to align the gyral and sulcal landmarks between the source and the target brain. Jacobian distortion is a measure of global brain deformation, and has been used to characterize the cortex for analyzing a number of neurological disorders [13].

2.2 Automated FCD Lesion Detection

The machine learning task is to develop a classifier that can distinguish between normal and abnormal cortical tissue using the patient's MRI data. SBM has been used in conjunction with machine learning and statistical techniques to identify lesions in FCD patients. Besson et al. [4] use texture, GWC and a number of morphological features including cortical thickness to represent each vertex on the reconstructed cortical surface. They then train a neural network to classify each vertex as being normal or lesional. Similarly, Thesen et al. [35] use a univariate z-score based thresholding approach on registered SBM data to classify each vertex as being lesional or normal. Recently, Hong et al. [17] developed a two-stage Fisher linear discriminant analysis (LDA) [5] classifier to detect FCD lesions in MRI-negative patients. Initially they train a vertex-level classifier that classifies each vertex on the reconstructed cortical surface as being lesional or non-lesional for both controls and patients. These detections are further refined using another LDA classifier that is trained to distinguish between actual FCD lesions (detections made inside the manually refined resection zones of patients) and false lesional detections made on controls.

One of the major confounding factors inhibiting the development of an effective classifier for detecting FCD lesions is *inter-patient variability*. The morphology of the human brain such as its thickness, curvature and the overall structure in general are affected by different factors such as age, gender, handedness, etc. [30]. This causes a co-variate shift in data as the data from different patients is pooled to learn a common classifier. Similarly, the distribution of pathological features that define an FCD lesion differs across FCD subtypes. For example in addition to causing other morphological abnormalities, FCD type I lesions appear on MRI as abnormally thin regions of cortex, while FCD type II is characterized by abnormally thick regions. This heterogeneity of feature distributions that define the target concept (FCD lesion) for the learner must be taken into account to develop an effective supervised lesion detection scheme.

To counter the effects of differences in feature distributions across FCD subtypes, Hong et al. deal only with FCD type-II [17], while Ahmed et al. [1], stratify the training data into thick and thin lesions based on manual inspection of data, while other schemes bypass this training bias by posing lesion detection as an outlier/anomaly detection problem [35, 2]. However, none of the lesion detection schemes cited previously explicitly address the co-variate shift arising from inter-patient variability. To overcome this co-variate shift in the underlying data distributions, we train an ensemble of classifiers on a training set consisting of an equal number of patients from each of the three FCD subtypes, and control data taken from fifty neurotypical controls. In order to discover subgroups in data, that align with meaningful combinations of patient and FCD subtype characteristics we use a Dirichlet process based mixture of RBMs.

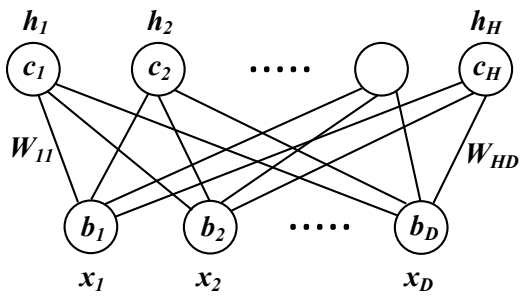


Figure 1: An RBM with D visible units and H hidden units

3. RESTRICTED BOLTZMANN MACHINES

Before describing our method in detail, we first review RBMs and DP mixture models.

3.1 Definition

A restricted Boltzmann machine (RBM, [32, 40]), illustrated in Figure 1, is an undirected graphic model that consists of two layers: a visible layer $\mathbf{x} = \{x_1, x_2, \ldots, x_D\}$, which represents the attributes of the input data, and a hidden layer $\mathbf{h} = \{h_1, h_2, \ldots, h_H\}$. Units across different layers are fully connected. However, connections within the same layer are restricted. The goal of an RBM is to model the distribution of the observations $\mathbf{x}$ with the help of hidden units $\mathbf{h}$.

We define the energy of an RBM as:

$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^\mathbf{T}\mathbf{W}\mathbf{x} - \mathbf{c}^\mathbf{T}\mathbf{x} - \mathbf{b}^\mathbf{T}\mathbf{h} \quad (1)$$

where $\mathbf{W} \in R^{H \times D}$ represents the weights connecting hidden and visible units and $\mathbf{b} \in R^D$, and $\mathbf{c} \in R^H$ are the offsets of the visible and hidden layers respectively. (1) directly leads to the following formulation for the probability density of $\mathbf{x}$:

$$p(\mathbf{x}) = e^{-F(\mathbf{x})}/\mathcal{Z} \quad (2)$$

where

$$F(\mathbf{x}) = \mathbf{c}^T\mathbf{x} + \sum_{j=1}^{H} \ln\left(1 + e^{(b_j + \mathbf{W}_j\mathbf{x})}\right) \quad (3)$$

$F(\mathbf{x})$ is often referred to as the free energy. H is the total number of hidden units and $\mathcal{Z} = \sum_{\mathbf{x}} e^{-F(\mathbf{x})}$ is the partition function. It can be shown that the conditional probability of h_i's given $\mathbf{x}$ is independent, i.e.,

$$p(\mathbf{h}|\mathbf{x}) = \prod_j p(h_j|\mathbf{x}) \quad (4)$$

and furthermore,

$$p(h_j = 1|\mathbf{x}) = \frac{1}{1 + exp(-(b_j + \mathbf{W}_j\mathbf{x}))} = \sigma(b_j + \mathbf{W}_j\mathbf{x}) \quad (5)$$

where $\sigma(\cdot)$ is the sigmoid function. Similarly, we have

$$p(\mathbf{x}|\mathbf{h}) = \prod_k p(x_k|\mathbf{h}) \tag{6}$$

$$p(x_k = 1|\mathbf{h}) = \frac{1}{1 + exp(-(c_k + \mathbf{h}^T \mathbf{W}_k))} = \sigma(c_k + \mathbf{h}^T \mathbf{W}_k) \tag{7}$$

Equations (4) - (7) allow us to make approximations to the inference algorithm described next.

3.2 Inference

Given the training data, the learning objective of an RBM is to adjust parameters $\{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$ such that the free energy defined in (3) is minimized, which is equivalent to maximizing $p(\mathbf{x})$. A standard approach is to use stochastic gradient descent to minimize the average negative log-likelihood $\frac{1}{T}\sum_{t=1}^{T} -\ln p(\mathbf{x}^t)$. Thus, we calculate the partial derivatives with respect to each parameter and set them to zero to obtain a set of update rules. A learning algorithm iteratively updates the parameters until convergence. Specifically,

$$\begin{aligned} -\frac{\partial \ln p(\mathbf{x})}{\partial \theta} &= \frac{\partial F(\mathbf{x})}{\partial \theta} + \frac{1}{\mathcal{Z}}\frac{\partial \mathcal{Z}}{\partial \theta} \\ &= \frac{\partial F(\mathbf{x})}{\partial \theta} + \frac{1}{\mathcal{Z}}\sum_{\mathbf{x}} e^{-F(\mathbf{x})}\frac{\partial(-F(\mathbf{x}))}{\partial \theta} \\ &= \frac{\partial F(\mathbf{x})}{\partial \theta} - \sum_{\mathbf{x}} p(\mathbf{x})\frac{\partial F(\mathbf{x})}{\partial \theta} \end{aligned} \tag{8}$$

where $\theta \in \{\mathbf{c}, \mathbf{b}, \mathbf{W}\}$. The computational difficulty in (8) is the second term which is an expectation over an exponential number of configurations of the input $\mathbf{x}$ under $p(\mathbf{x})$. Hinton introduced the Contrastive Divergence (CD) learning algorithm [15] which makes this computation tractable by estimating the expectation using a sample, $\tilde{\mathbf{x}}$ from the model. This sample is obtained using k-steps of Gibbs sampling. Thus, (8) can be simplified to:

$$-\frac{\partial \ln p(\mathbf{x})}{\partial \theta} = \frac{\partial F(\mathbf{x})}{\partial \theta} - \frac{\partial F(\tilde{\mathbf{x}})}{\partial \theta} \tag{9}$$

For RBMs, because the hidden and visible units are conditionally independent (Equation (4), (6)), we can perform block Gibbs sampling, i.e., we sample all hidden (visible) units simultaneously given the values of the visible (hidden) units. In particular, starting with a given visible observation $\mathbf{x}^{(n)}$, we have:

$$\mathbf{h}^{(n+1)} \sim \sigma(\mathbf{W}'\mathbf{x}^{(n)} + \mathbf{c})$$

$$\mathbf{x}^{(n+1)} \sim \sigma(\mathbf{W}\mathbf{h}^{(n+1)} + \mathbf{b})$$

Consequently, the Contrastive Divergence learning algorithm with k steps of Gibbs sampling (CD-k) can be summarized as follows:

1. For each training example $\mathbf{x}^{(t)}$
   (a) Starting at $\mathbf{x}^{(t)}$, generate a sample $\tilde{\mathbf{x}}$ using k steps of Gibbs sampling.
   (b) Update parameters

$$\begin{aligned} \mathbf{W} &\Longleftarrow \mathbf{W} + \alpha\left(h(\mathbf{x}^{(t)})\mathbf{x}^{(t)^T} - h(\tilde{\mathbf{x}})\tilde{\mathbf{x}}^T)\right) \\ \mathbf{b} &\Longleftarrow \mathbf{b} + \alpha\left(h(\mathbf{x}^{(t)}) - h(\tilde{\mathbf{x}}))\right) \\ \mathbf{c} &\Longleftarrow \mathbf{c} + \alpha\left(\mathbf{x}^{(t)} - \tilde{\mathbf{x}}\right)\right) \end{aligned}$$

   where $h(\mathbf{x}) = \sigma(\mathbf{b} + \mathbf{W}\mathbf{x})$ and α is the learning rate.

2. Go to Step 1 until the stopping criterion is met.

We set $k = 1$ in our model because it has been shown empirically that even when k is not large (e.g., $k = 1$), the CD-k algorithms often gives good results [9].

4. DIRICHLET PROCESS (DP) MIXTURE MODEL

Dirichlet process is a family of Bayesian nonparametric models in which the model representations grow as more data are observed [39, 38, 26]. In particular, DP used as a prior in a generative mixture model allows the number of mixing components to adapt to the individual dataset automatically. DP can be interpreted as an extension to the traditional generative model with an arbitrary (infinite) number of mixing components.

Formally, a Dirichlet process is an infinite dimensional discrete distribution with two parameters α and H denoted as:

$$G \sim DP(\alpha, H)$$

where H is the base distribution and scalar α is the strength parameter. H serves as the mean of G and α controls the convergence of G towards H. A Dirichlet process can be constructed using the stick-breaking process [31] as follows:

$$\begin{aligned} \theta_k^* &\sim H & v_k &\sim Beta(1, \alpha) \\ \pi_k^* &= v_k \prod_{j=1}^{k-1}(1 - v_j) & G &= \sum_{k=1}^{\infty} \pi_k^* \delta(\theta_k^*) \end{aligned} \tag{10}$$

where $k = 1, 2, \ldots$. and δ is the Dirac delta function.

A DP mixture model uses $G(\alpha, H)$ as the prior under the Bayesian framework. The entire dataset is modeled as a mixture of components and each component is parameterized by a random draw (θ) from G. Each data observation belongs to one of the components and is modeled as a function of the parameter of its component, i.e., $f_i(\theta_i)$. Specifically,

$$\begin{aligned} G|\alpha, H &\sim DP(\alpha, H) & \theta_i|G &\sim G \\ x_i|\theta_i &\sim f_i(\theta_i) \end{aligned}$$

Consider drawing N samples of θ_i ($i = 1, 2, \ldots, N$) from G. Because G is a discrete distribution, the probability at any given point in the probability space can be non-zero. This implies that the values of the θ_i's will repeat with a positive probability. Hence, these θ_i's exhibit clustering behavior (Polya Urn Scheme)[33]. Given the first N samples of θ_i from G, we assume they have produced a set of k distinct values:

$$\Theta^* = \{\theta_1^*, \theta_2^*, \ldots, \theta_k^*\} \text{ where } k < N.$$

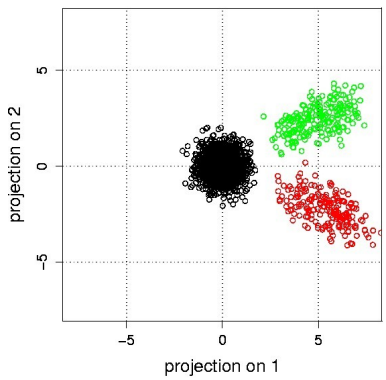


Figure 2: Projection of Clusters

It can be shown that the next new sample θ_{N+1} can be either a new value drawn from base distribution H with probability $\propto \alpha$ or can be taken from one of the existing members from Θ^* with probability $\propto c_i$, where c_i is the number of times θ_i^* has been repeated. Specifically,

$$\theta_{N+1}|\theta_1, \theta_2, \ldots, \theta_N \sim \frac{\alpha}{\alpha+N}H + \sum_{i=1}^{n} \frac{c_i}{\alpha+N}\delta_{\theta_i} \qquad (11)$$

where δ_{θ_i} denotes the distribution concentrated at a single point θ_i.

Equation (11) illustrates two important properties of a DP. First, the concentration parameter α controls the number of distinct values of θ_i's, i.e., the number of mixing components. Second, DP exhibits a "rich get richer" property: the more frequently a θ_i^* has been adopted (i.e., the larger the c_i), the more likely it will be chosen again as the next θ_i value.

There are various techniques such as Markov chain Monte Carlo (MCMC) sampling[24] and variational inference[7] to conduct inference for a non-parametric model under the Bayesian framework. The variational approach can be more advantageous due to its scalability and guaranteed convergence. In this paper, we adopt the mean-field variational approach outlined in [7].

5. RBM-DPM MODEL

In this section, we describe how to combine the RBM and DPM algorithms for our classification task, including a modified weighted RBM training algorithm.

5.1 Integrating RBM and DPM

The inference of a typical mixture (finite or infinite) of RBMs is intractable due to the difficulty of estimating the partition function $\mathcal{Z}$ in density function $p(\mathbf{x})$ defined in Equation (2). We propose to take a two-stage approach. We first employ a DPM model to partition the data into k clusters. We then apply a modified version of RBM training algorithm to each non-empty cluster using the weights obtained from the DPM model.

For stage one, we note that it is the top layer hidden units (i.e., the non-linear features extracted by the RBM) that serve as the input to our classifier. Thus, directly clustering in the input $\mathbf{x}$ may not be effective for a task for which the inputs are the $\mathbf{h}$'s. For example, in image processing, we can interpret the $\mathbf{h}$'s as a projection of $\mathbf{x}$'s on some other dimension. In Figure 2, the three clusters in the original input space are no longer valid when projected to either of the two dimensions. Input data, therefore, should be partitioned according to the characteristics of the projections (i.e., the top layer hidden units) rather than the original observations.

We propose applying an infinite mixture model to cluster the data at the top layer of the neural network, i.e., the $\mathbf{h}$'s in an RBM, and then learn k classifiers for those partitions produced by the clustering algorithm. To predict the label for a new instance, we will feed the instance to each classifier and take a majority vote from their predictions. Our RBM-DPM model can be outlined as follows:

RBM-DPM Classification Model

INPUT: Instances $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
Labels $\mathbf{L} = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n\}$

1. Train an RBM R_0 $(\mathbf{W_0}, \mathbf{b_0}, \mathbf{c_0})$ using the entire input dataset $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$.
2. Compute the set of hidden units values, $\mathbf{h} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$, where $\mathbf{h_i} = \mathbf{W}\mathbf{x_i}+\mathbf{b}$ and $\mathbf{W}, \mathbf{b}$ are learned parameters from R_0.
3. Train a DPM mixture model (see Section 5.2) on $\mathbf{h} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$, producing k non-empty clusters.
4. Train k weighted RBMs (see Section 5.3) $R_1, R_2, \ldots R_k$ for each non-empty cluster from Step 3.
5. Transform each R_i $(i = 1, 2, \ldots, k)$ into a classifier by augmenting an output layer of units. Adjust parameters $(\mathbf{W}_i, \mathbf{b}_i, \mathbf{c}_i)$ using labels $\mathbf{L}$ and the backpropagation algorithm.

OUTPUT: Classifiers $R_1, R_2, \ldots, R_k$ on input domain $\mathbf{x}$.

5.2 Mixture of Hidden Units

In this section, we give the formulation of the DP mixture model of the hidden units $\mathbf{h} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$. It corresponds to Step 3 of the RBM-DPM model. We assume instances in the k^{th} cluster follow a multivariate Gaussian distribution $\mathcal{N}(\mu_k, \Sigma_k)$. Our model employs the latent variables $\mathbf{z} = \{z_1, z_2, \ldots\}$ (z_i's are the rows of the indicator matrix $\mathbf{z}$ mentioned previously), $\mathbf{v} = \{v_1, v_2, \ldots, \}$ (v_i's are the stick-breaking proportions [31] in a DP) and $\Theta = \{\mu_1, \mu_2, \ldots, \Sigma_1, \Sigma_2, \ldots\}$. Hyper-parameters are α, μ_0 and Σ_0.

The joint distribution $p(\mathbf{h}, \mathbf{z}, \mathbf{v}, \Theta|\mathbf{x})$ can be factored as follows:

$$p(\mathbf{h}, \mathbf{z}, \mathbf{v}, \Theta|\mathbf{x}) = p(\mathbf{h}|\mathbf{z}, \mathbf{\Theta}, \mathbf{x})p(\Theta)p(\mathbf{z}|\mathbf{v})p(\mathbf{v}|\alpha) \qquad (12)$$

where

$$p(\mathbf{v}|\alpha) = \prod_{k=1}^{\infty} \text{Beta}(v_k|1,\alpha)$$

$$p(\mathbf{z}|\mathbf{v}) = \prod_{n=1}^{N}\prod_{k=1}^{\infty}\left(v_k\prod_{j=1}^{k-1}(1-v_j)\right)^{z_{n,k}}$$

$$p(\Theta) = \prod_{k=1}^{\infty}\mathcal{N}(\Theta_k|\mu_0,\Sigma_0)$$

$$p(\mathbf{h}|\mathbf{z},\Theta,\mathbf{x}) = \prod_{n=1}^{N}\prod_{k=1}^{\infty}\mathcal{N}(\mathbf{h}_n|\mu_k,\Sigma_k)^{z_{n,k}}$$

This is a standard DPM of Gaussian distributions, for which we use the variational inference method [6, 7] to estimate the parameters. The inference algorithm iteratively computes the values of latent variables until convergence. In particular, the mixture model produces a soft mixture of $\mathbf{h}$'s as follows:

$$\begin{bmatrix} z_{1,1}\cdot\mathbf{h}_1 & z_{1,2}\cdot\mathbf{h}_1 & \dots & z_{1,k}\cdot\mathbf{h}_1 & \dots \\ z_{2,1}\cdot\mathbf{h}_2 & z_{2,2}\cdot\mathbf{h}_2 & \dots & z_{2,k}\cdot\mathbf{h}_2 & \dots \\ \vdots & \vdots & \ddots & \vdots & \\ z_{n,1}\cdot\mathbf{h}_n & z_{n,2}\cdot\mathbf{h}_n & \dots & z_{n,k}\cdot\mathbf{h}_n & \dots \end{bmatrix}$$

where each column is a cluster and $\sum_j z_{i,j} = 1$ for $i = 1, 2, \dots, n$.

We next proceed to the second stage, i.e., Step 4 of the RBM-DPM model, and train a weighted RBM (described below) for each non-empty cluster. The weights for cluster k are defined by the $z_{i,k}$'s ($i = 1, 2, \dots, n$).

5.3 Weighted RBM

The mixing proportions for the i^{th} cluster are:

$$\begin{bmatrix} z_{1,i} \\ z_{2,i} \\ \vdots \\ z_{n,i} \end{bmatrix}$$

Because the $\mathbf{h}_i$'s are obtained from the corresponding $\mathbf{x}_i$'s, we modify the energy function using weighted $\mathbf{x}_i$'s:

$$F(\mathbf{x}) = \mathbf{c}^T D\mathbf{x} + \sum_{j=1}^{H}\ln\left(1+e^{(b_j+\mathbf{W}_j D\mathbf{x})}\right)$$

$$\text{where} \quad D = \begin{bmatrix} z_{1,i} & & & \\ & z_{2,i} & & \\ & & \ddots & \\ & & & z_{n,i} \end{bmatrix}$$

Consequently, we can modify the CD-k learning algorithm as follows:

Table 1: Epilepsy Type and Data Instances Contributed from Each Patient

Patient	Type	Positive Instances	Negative Instances
NY143	2	629	629 * 50
NY148	1	6,569	6,569 * 50
NY149	2	7,035	7,035 * 50
NY159	1	5,111	5,111 * 50
NY186	3	5,869	5,869 * 50
NY294	3	14,107	14,107 * 50
NY226	1	6,485	6,485 * 50
NY255	2	14,394	14,394 * 50
NY259	1	6,792	6,792 * 50
NY315	2	3,434	3,434 * 50
NY338	3	9,046	9,064 * 50
NY343	3	10,463	10,463 * 50
NY351	1	3,522	3,522 * 50
NY371	2	6,915	6,915 * 50
NY394	2	14,197	14,197 * 50
NY46	1	18,972	18,972 * 50
NY67	3	14,932	14,932 * 50
NY72	3	15,448	15,448 * 50
Total		163,920	8,196,000

Weighted CD-k Learning Algorithm

1. For each training example $\mathbf{x}^{(t)}$
   (a) Starting at $\mathbf{x}^{(t)}$, generate a sample $\tilde{\mathbf{x}}$ using k steps of Gibbs sampling.
   (b) Update parameters

$$\mathbf{W} \Longleftarrow \mathbf{W} + \alpha D\left(h(\mathbf{x}^{(t)})\mathbf{x}^{(t)^T} - h(\tilde{\mathbf{x}})\tilde{\mathbf{x}}^T)\right)$$

$$\mathbf{b} \Longleftarrow \mathbf{b} + \alpha D\left(h(\mathbf{x}^{(t)}) - h(\tilde{\mathbf{x}}))\right)$$

$$\mathbf{c} \Longleftarrow \mathbf{c} + \alpha D\left(\mathbf{x}^{(t)} - \tilde{\mathbf{x}})\right)$$

   where $h(\mathbf{x}) = \sigma(\mathbf{b} + \mathbf{W}\mathbf{x})$, α is the learning rate and D is the weight matrix.

2. Go to Step 1 until stopping criterion is met.

6. EXPERIMENTAL RESULTS

In this section we describe in detail the construction of our dataset, the machine learning techniques used to address domain-specific issues related to our task, and the performance evaluation of our proposed RBM-DPM model.

6.1 Patient and Data Description

As described in Section 1, we had access to the MRI data of MRI-negative patients from NYU's Comprehensive

Table 2: RBM-DPM Model Compared to LR and RBM Models with Different Ensemble Thresholds

Patient	Threshold = 90						Threshold = 95					
	LR		RBM		RBM+DPM		LR		RBM		RBM+DPM	
	-		# Classifiers*: 18		# Classifiers: 25		-		# Classifiers: 12		# Classifiers: 17	
	Detected	TNR	Detected	TNR	Detected	TNR	Detected	TNR	Detected	TNR	Detected	TNR
NY226	Y	96.3	Y	99.0	Y	99.6	Y	98.7	N	98.3	Y	99.9
NY255	Y	96.8	Y	97.5	Y	99.6	Y	98.6	Y	99.3	Y	99.9
NY259	N	96.3	Y	98.3	Y	99.7	N	99.4	N	97.6	Y	99.9
NY315	N	97.7	Y	97.9	N	99.6	N	98.5	N	99.0	N	99.9
NY338	Y	98.5	Y	97.6	Y	99.4	N	99.8	Y	99.3	Y	99.8
NY343	Y	97.1	Y	97.6	Y	99.6	Y	99.1	Y	98.4	Y	99.9
NY351	N	97.9	N	97.2	N	99.7	N	98.3	N	98.3	N	99.9
NY371	Y	97.5	Y	97.2	N	99.6	Y	99.5	Y	98.1	N	99.9
NY394	N	98.1	N	97.5	N	99.7	N	99.6	N	98.1	N	99.9
NY46	Y	97.4	Y	98.7	Y	99.7	Y	99.2	Y	97.9	Y	99.9
NY67	Y	97.3	Y	98.6	Y	99.6	Y	99.3	Y	99.3	Y	99.8
NY72	N	98.1	N	99.8	N	99.8	N	99.7	N	98.4	N	99.9
Mean	58%	97.4	75%	98.1	58%	99.6	50%	99.1	50%	98.6	58%	99.9

*number of retained classifiers after the selective ensemble described in Section 6.3.

Epilepsy Center who underwent resective surgery and were completely seizure free after surgery. Furthermore, their resected cortical tissue was histopathologically verified to contain FCD. These patients are further categorized into three subtypes (type I, II and III) [8]. Six representative patients were selected from each subtype group resulting in a total of eighteen patients available to our research. This might seem like a small collection of patients, however it should be noted that only a few MRI-negative patients proceed to surgical resection and out of these few only a third achieve complete post-surgical seizure freedom. *Indeed, the six type II patients in our collection represent the entire population of FCD type-II MRI-negative patients treated at NYU during the past three years.*

In terms of the actual data instances, each patient contributes a different number of positive instances to our dataset, depending on the size of his/her resection area. All the vertices within the resected region are labeled as positive (lesional, label = 1). The negative (non-lesional, label = 0) instances for our dataset are extracted from the MRI scans of fifty neurotypical healthy controls who underwent the same MRI protocol. In particular, for each patient, we extract data from each healthy image from the same location as the patient's resection region. As a result, if a patient contributes n lesional samples to our dataset, we will have fifty corresponding sets of non-lesional samples with n instances in each set (i.e., a total of $50 * n$ instances). Table 1 presents the subgroup type, total number of positive and negative instances associated with each patient. We have a total of 163,920 and 8,196,000 positive and negative instances respectively. Note that the negative instances in our dataset are taken from the healthy controls instead of from non-lesional areas outside the patient resection zones. This approach encourages our model to learn a more accurate representation of normal human brains, which is essential to our task of abnormality detection.

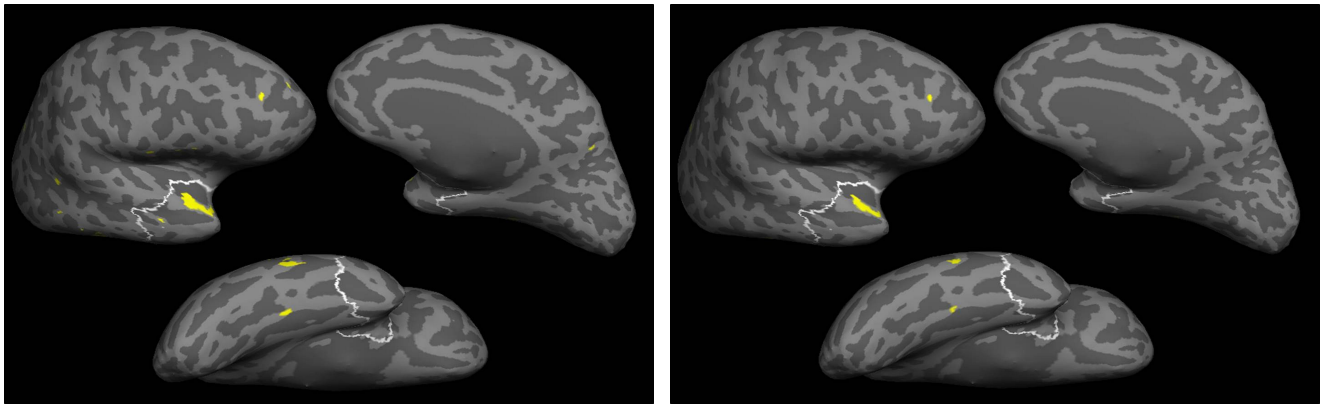
6.2 Constructing Training Set

The majority of the FCD patients in our dataset have temporal lobe resections, which is the most prevalent localization of FCD in adults [19]. Training on all patients would therefore be biased toward a specific cortical region limiting its generalization to differentiate between lesional and non-lesional vertices in other cortical regions. Training on all patients (and performing evaluation via leave-one-patient out cross validation) would result in low accuracy in regions other than the temporal lobe. Furthermore, it is desirable to have a balanced training set of patients from the three different FCD subtypes to give a good distribution over different FCD lesion types. Under these two constraints, we selected two patients from each FCD subtype (i.e., six patients in total) as our training patients such that their resected regions optimize coverage of the cortex beyond the temporal lobe [1].

6.3 Selective Ensemble of Classifiers

There is a fifty to one ratio between negative and positive instances in our data. In order to overcome this class imbalance, we apply bagging with under-sampling [36]. Each bag includes all positive instances from our training patients (i.e., all minority instances). At the same time, we randomly pick one control for each patient and include all the corresponding negative instances from the selected control.

[1]In experiments, not reported in this paper due to space, we later verified our assumptions that training on a distribution skewed toward lesions in the temporal lobe did indeed lead to lower performance than a more balanced dataset. We omit these results due to space limitations.



(a) Threshold=90 (b) Threshold=95

Figure 3: Detection Results for NY255 using the proposed scheme, at two different thresholds of bag selection, plotted on the inflated cortical surface. The detected clusters are depicted as the solid yellow regions, while the white outlined region represents the actual resected region.

We repeat this process fifty times and create fifty balanced training sets. We learn a classifier for each training set resulting in fifty independent classifiers. A standard ensemble method in machine learning will perform a majority vote among all fifty classifiers when making a prediction. In our case, however, we are biased towards the learners that boast high TNRs on the training data. This is because, although we are aiming at a high lesion detection rate, our detection is only meaningful if we don't sacrifice the performance on non-lesional areas. To address this preference issue specific to our task, we establish an *Ensemble Threshold* and discard classifiers whose TNRs fall below this threshold *on the training data*. The goal is to weed out classifiers that fail to capture the characteristics of healthy cortical tissue.

6.4 Evaluation Method

We have twelve test patients (excluding the six training patients) and we present our model's performance on each test patient's entire brain i.e., vertices from both inside and outside the resected region.We measure the performance on the lesional and non-lesional instances separately to ensure that we are not only detecting the lesions, but also are not misclassifying normal cortical vertices. As discussed in Section 1, we use an indication flag to signal a successful detection within the resection zone and a TNR to measure our performance on non-lesional instances. Note that the benign instances of the patients are not part of our dataset constructed in Section 6.1. Thus, a good performance on these instances (i.e., a high TNR) demonstrates the efficacy of our model in recognizing new healthy brain structures.

We compare the performance of our RBM-DPM model to two baseline models. The first is our recently reported logistic regression (LR) based approach [1] which deals exclusively with MRI-negative patients. This model has been in use at NYU's Comprehensive Epilepsy Center since 2013 and entails a number of pre-processing steps which include manual reduction of the resected region, stratifying the data based on sulcal depth, and a post-processing step that manually discards all detected clusters that fall below a surface area of $50mm^2$. The performance of this model reported in Table 2 which includes all the pre-processing and post-processing steps, and thus represents this method's best performance.

For the second baseline model, we choose to train an RBM over the entire dataset without applying the DPM clustering algorithm to the data. The purpose of the comparison is to verify our conjecture that DPM is able to capture the interpersonal variations arising from different morphologies among the patients.

6.5 Discussion

Table 2 presents the main results of our model on the twelve test patients. We experimented with two thresholds (90 and 95, indicated by the first row) for retaining the classifiers (see details in Section 6.3). Lower thresholds are not interesting because they lead to lower TNR. For each threshold value, we compare the performance of our model to the two baseline models. The third row shows the number of classifiers retained in each case. Under columns "Detected", a value 'Y' indicates a successful detection. The corresponding entry in the last row shows the successful detection rate out of the twelve patients in each model.

We have developed our approach keeping in mind its final use as a focus of attention mechanism for neuroradiologists. The detections made using our model would be used to inform effective electrode placement in iEEG, and constitute a source of secondary evidence for determining the resection target with an ultimate goal of reducing the resection region which leads to a safer and effective surgery. Thus, there are two performance metrics that need to be analyzed: i) performance on the non-lesional regions (i.e., TNR) and ii) detection rate on the lesional regions.

- A high true negative rate is essential in defining success in our task as it alludes to the number of false detections that a neuroradiologist needs to inspect. Indeed, a model with a TNR rate below 99% is impractical because there will be too many false positive regions. One way to increase the TNR rate is to raise the *Ensemble Threshold* to a more stringent value (e.g., from 90 to 95). As we observe in Table 2, the TNR rates improve across all models as we move the threshold from 90 to 95.

- The RBM-DPM approach dominates both RBM and LR models by delivering the highest TNR rates (99.6% and 99.9%) in both threshold settings. Although the RBM model has a higher detection rate at threshold 90, its corresponding TNR rate (98.1%) is too low to be effective for the neuroradiologists. The LR model achieves satisfactory TNR performance at threshold 95, which is the setting that is currently used by NYU's Comprehensive Epilepsy Center.

- Detection rate is defined as the number of patients for whom an abnormal region is detected within their resected region. Although the detection rates presented in Table 2 seem low (50% to 75%), it is worth noting that the detections are made from MRI images of MRI-negative patients, which means the doctors have no identification of any abnormality in these MRI images.

- We observe from Table 2, that RBM-DPM model delivers a more stable performance compared to the other two models at the two threshold settings. In particular, RBM-DPM maintained a detection rate of 58% (7/12 patients) when the threshold changed from 90 to a more stringent value 95. On the other hand, both RBM and LR models sufferred a drop in their detection rates in the process. For the RBM model, the detection rate changed from 75% to 50% and for the LR model, the change was from 58% to 50%. This stable performance coupled with its near perfect TNR accuracies makes RBM-DPM a desirable tool in the pre-surgical evaluation process.

- Even though resective surgery is a viable option for FCD patients, it remains underutilized as most practitioners are unwilling to perform surgery in the absence of a visually detected lesion. This limits the number of available patients whose data can be used to build automated lesion detection models. Our model has the potential to increase the number of patients who are referred to surgery by locating the lesion during the pre-surgical evaluation process. Although the current sample may seem small, the results are significant since a board of experienced neuroradiologists failed to locate the lesion for all these patients, and our approach is able to detect the lesion in 58% of the patients with a near perfect TNR.

Figures 3(a) and 3(b) plot the detection results of RBM-DPM model on a test patient for the two different threshold values using an inflated model of their cortical surface. The yellow areas are detected lesional regions from the model and the white outlined region indicates the actual resected area. Thus, yellow clusters within the resected zone are correct detections, while those outside the zone are false positives. As explained in Section 1, the resection zones are determined in a "generous" manner for MRI-negative patients to maximize the chances of a seizure free outcome. Indeed, we observe in Figures 3(a) and 3(b) that the detected lesional regions are considerably smaller than the size of the actual resection. Thus, the results of our model can be used to guide electrode placement during a patient's iEEG evaluation and obtain a potentially refined resection target, which leads to reduced chances of removing healthy cortical tissue.

7. CONCLUSION

In this paper, we proposed a non-parametric approach to detect MRI-elusive epilepsy lesions using restricted Boltzmann machines (RBMs). In particular, we transform 3D-MRI images of human brains into a standard 2D surface using the Surface-Based Morphometry methodology and extract five features that characterize human cortical surfaces. Our model addresses both issues of limited available features and inter-patient variabilities in the input data. For the former, we used an RBM as a pre-training step. For the latter, we applied a Dirichlet process based clustering algorithm and estimated its parameters via variational inference. To accomplish our classification task, we collect multiple classifiers by training an augmented RBM for each non-empty component from the clustering algorithm and take a majority vote among all classifiers while making a prediction. We evaluated our model on brain images of twelve MRI-negative patients Our model correctly detected abnormal regions within the resected areas in 58% of the patients, with 99.9% accuracy of correctly classifying the non-lesional vertices. Based on these findings, we are evaluating a replacement of our current LR model with our new approach in the clinical treatment for epilepsy patients at NYU's Comprehensive Epilepsy Center.

8. ACKNOWLEDGMENTS

This work is partially supported by NSF IIS-1546428. We would like to thank Hugh Wang at the neurocognitive laboratory, New York University for his help with mapping the resection masks.

9. ADDITIONAL AUTHORS

Additional authors: Ruben Kuzniekcy (New York University, email: `ruben.kuzniecky@nyumc.edu`) and Orrin Devinsky (New York University, email: `od4@med.nyu.edu`).

10. REFERENCES

[1] B. Ahmed, C. E. Brodley, K. E. Blackmon, R. Kuzniecky, G. Barash, C. Carlson, B. T. Quinn, W. Doyle, J. French, O. Devinsky, and T. Thesen. Cortical feature analysis and machine learning improves detection of mri-negative focal cortical dysplasia. *Epilepsy & Behavior*, 48:21 – 28, 2015.

[2] B. Ahmed, T. Thesen, K. Blackmon, Y. Zhao, O. Devinsky, R. Kuzniecky, and C. Brodley. Hierarchical conditional random fields for outlier detection: An application to detecting epileptogenic cortical malformations. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1080–1088, 2014.

[3] A. Bernasconi, N. Bernasconi, B. Bernhardt, and D. Schrader. Advances in mri for 'cryptogenic' epilepsies. *Nat Rev Neurol.*, 7(2):99–108, 2011.

[4] P. Besson, N. Bernasconi, O. Colliot, et al. Surface-based texture and morphological analysis detects subtle cortical dysplasia. In *MICCAI*, pages 645–652, 2008.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[6] D. M. Blei and M. I. Jordan. Variational methods for the dirichlet process. *Proceedings of the twenty-first*

international conference on Machine learning, page 12, 2004.

[7] D. M. Blei and M. I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1.1:121–143, 2006.

[8] I. Blumcke, M. Thom, E. Aronica, et al. The clinicopathologic spectrum of focal cortical dysplasias: A consensus classification proposed by an ad hoc task force of the ILAE diagnostic methods commission. *Epilepsia*, 52(1):158–174, 2011.

[9] M. A. Carreira-Perpinan and G. Hinton. On contrastive divergence learning. *AISTATS*, 10:33–40, 2005.

[10] A. Dale, B. Fischl, and M. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999.

[11] B. Fischl and A. Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *PNAS*, 97(20):11050–11055, 2000.

[12] B. Fischl, M. Sereno, and A. Dale. Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, 1999.

[13] C. D. Good, J. Ashburner, and R. Frackowiak. Computational neuroanatomy: new perspectives for neuroradiology. *Revue Neurologique*, 157:685–700, 2001.

[14] W. A. Hauser and D. C. Hesdorffer. *Epilepsy: frequency, causes and consequences*. Epilepsy Foundation of America, 1990.

[15] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14.8:1771–1800, 2002.

[16] G. E. Hinton. Boltzmann machine. *Scholarpedia*, 2(5):1668, 2007.

[17] S. J. Hong, H. Kim, D. Schrader, N. Bernasconi, B. C. Bernhardt, and A. Bernasconi. Automated detection of cortical dysplasia type II in MRI-negative epilepsy. *Neurology*, 83(1):48–55, 2014.

[18] D. Killer and N. Friedman. Probabilistic graphical models: Principles and techniques. *MIT Press*, 2009.

[19] P. Krsek, B. Maton, B. Korman, E. Pacheco-Jacome, P. Jayakar, C. Dunoyer, et al. Different features of histopathological subtypes of pediatric focal cortical dysplasia. *Annals of Neurology*, 63(6):758–769, 2008.

[20] R. I. Kuzniecky and A. Barkovich. Malformations of cortical development and epilepsy. *Brain and Development*, 23(1):2 – 11, 2001.

[21] P. Kwan and M. J. Brodie. Early identification of refractory epilepsy. *New England Journal Of Medicine*, 342(5):314–319, 2000.

[22] J. T. Lerner et al. Assessment and surgical outcomes for mild type I and severe type II cortical dysplasia: a critical review and the UCLA experience. *Epilepsia*, 50(6):1310–1335, 2009.

[23] V. Nair and G. E. Hinton. Implicit mixtures of restricted boltzmann machines. *dvances in neural information processing systems*, pages 1145–1152, 2009.

[24] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9.2:249–265, 2000.

[25] C. Nordahl, D. Dierker, I. Mostafavi, et al. Cortical folding abnormalities in autism revealed by surface-based morphometry. *J Neurosci.*, 27(43):11725–11735, 2007.

[26] J. Paisley, C. Wang, D. Blei, and M. I. Jordan. A nested hdp for hierarchical topic models. *arXiv preprint arXiv:1301.3570*, 2013.

[27] R. P. R, B. Fischl, V. C. V, N. Makris, and P. E. Grant. A methodology for analyzing curvature in the developing brain from preterm to adult. *International Journal of Imaging Systems and Technology*, 18(1):42–68, 2008.

[28] L. Rimol, R. NesvÃëg, D. Hagler Jr., et al. Cortical volume, surface area, and thickness in schizophrenia and bipolar disorder. *Biological Psychiatry*, 71(6):552–560, 2012.

[29] D. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1:318–362, 1986.

[30] D. H. Salat, R. L. Buckner, A. Snyder, et al. Thinning of the cerebral cortex in aging. *Cerebral Cortex*, 14(7):721–730, 2004.

[31] J. Sethuraman. A constructive definition of dirichlet priors. *FLORIDA STATE UNIV TALLAHASSEE DEPT OF STATISTICS*, No. FSU-TR-M-843, 1991.

[32] P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. *Parallel Distributed Computing: Explorations in the Microstructure of Cognition*, 1, 1986.

[33] Y. W. Teh. Dirichlet process. *Encyclopedia of machine learning. Springer US*, pages 280–287, 2010.

[34] J. F. Tellez-Zenteno, R. Dhar, and S. Wiebe. Long-term seizure outcomes following epilepsy surgery: a systematic review and meta-analysis. *Brain*, 128(5):1188–1198, 2005.

[35] T. Thesen, B. Quinn, C. Carlson, et al. Detection of epileptogenic cortical malformations with surface-based MRI morphometry. *PLoS ONE*, 6(2):e16430, 2011.

[36] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos. Class imbalance, redux. *In Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 754–763, 2011.

[37] Z. I. Wang, A. V. Alexopoulos, S. E. Jones, Z. Jaisani, I. M. Najm, and R. A. Prayson. The pathology of magnetic-resonance-imaging-negative epilepsy. *Mod Pathol*, 26(8):1051–1058, 2013.

[38] E. P. Xing, M. I. Jordan, and R. Sharan. Bayesian haplotype inference via the dirichlet process. *Journal of Computational Biology*, 14.3:267–284, 2007.

[39] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with dirichlet process priors. *The Journal of Machine Learning Research*, 8:35–63, 2007.

[40] F. Y. and D. Haussler. Unsupervised learning of distributions on binary vectors using 2-layer networks. *NIPS*, pages 912–919, 1992.