Semi-Supervised Anomaly Detection for EEG Waveforms Using Deep Belief Nets

Drausin Wulsin, Justin Blanco
Department of Bioengineering
University of Pennsylvania
Philadelphia, PA, USA
{wulsin, blancoj}@seas.upenn.edu

Ram Mani, Brian Litt
Department of Neurology
University of Pennsylvania
Philadelphia, PA, USA
ram.mani@uphs.upenn.edu, littb@mail.med.upenn.edu

***Abstract*—Clinical electroencephalography (EEG) is routinely used to monitor brain function in critically ill patients, and specific EEG waveforms are recognized by clinicians as signatures of abnormal brain. These pathologic EEG waveforms, once detected, often necessitate accute clinincal interventions, but these events are typically rare, highly variable between patients, and often hard to separate from background, making them difficult to reliably detect. We show that Deep Belief Nets (DBNs), a type of multi-layer generative neural network, can be used effectively for such EEG anomaly detection. We compare this technique to the state-of-the-art, a one-class Support Vector Machine (SVM), showing that the DBN outperforms the SVM by the F_1 measure for our EEG dataset. We also show how the outputs of a DBN-based detector can be used to aid visualization of anomalies in large EEG data sets and propose a method for using DBNs to gain insight into which features of signals are characteristically anomalous. These findings show that Deep Belief Nets can facilitate human review of large amounts of clinical EEG as well as mining new EEG features that may be indicators of unusual activity.**

***Keywords*-anomaly, detection, outlier, novelty, Deep Belief Nets, EEG, data mining**

I. Introduction

A. Clinical Problem

Neurophysiological clinical monitoring routinely uses electroencephalography (EEG) to assess brain function and state in critically ill patients. Such monitoring often produces very large amounts of data[1] that is impractical for physicians to manually review. Even in patients with severe brain pathology, a large majority of the waveforms are still considered background or normal. Such patients have occasional discharges that are unlike the background signal and of diagnostic interest. The variability and poorly defined nature of these signals as well as that of background EEG may often make simple linear detection methods untenable [1]. Robust, nonlinear automated detection strategies of these unusual waveforms are therefore of great clinical interest.

In order to better understand the highly variable anomalous signals, human learning from the detector is also of great clinical interest. An ideal detector would thus both find anomalies and allow us to visualize what separates them from background.

[1]A typical patient receiving continuous EEG monitoring on 24 channels sampled at 256 Hz can generate roughly 1 GB of data per day. Intracranial EEG, with many more channels and higher sampling rates can produce orders of magnitude more data (e.g. 500 GB) per day.

B. DBN Anomaly Detection

Detection of unusual signals or anomalies (also called outliers or novelties) is a mature and active field of machine learning research [2]. While many types of anomaly detection methods exist, we used Deep Belief Nets (DBNs), a type of multi-layer generative neural network, for their ability to model numerous types of high-dimensional data and visualize high-level learned features [3], [4]. A thorough description of DBN varieties and their training is available elsewhere [4], [5]. We consider a DBN composed of logistic Restricted Boltzmann Machines (RBMs) with symmetric weights $\mathbf{W}$ between binary visible units $\mathbf{v}$ and binary hidden units $\mathbf{h}$ as well as biases $\mathbf{b}$ and $\mathbf{c}$ to the hidden layer and visible layer. An RBM has a joint distribution

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{z} e^{-\mathbf{h}^T \mathbf{W} \mathbf{v} - \mathbf{b}^T \mathbf{h} - \mathbf{c}^T \mathbf{v}} \tag{1}$$

where z is a normalization constant and thus has an energy function

$$E(\mathbf{v}, \mathbf{h}) = \mathbf{h}^T \mathbf{W} \mathbf{v} + \mathbf{b}^T \mathbf{h} + \mathbf{c}^T \mathbf{v} \tag{2}$$

The binary units of the hidden layer are Bernoulli random variables, where each hidden unit h_j is activated, here with the sigmoid function, based on each visible unit v_i with probability

$$P(h_j = 1) = \text{Sigm}(b_j + \sum_i v_i W_{ij}) \tag{3}$$

Calculating the gradient of the log likelihood of $\mathbf{v}$ is intractable and so contrastive divergence with k iterations (often $k = 1$) [6] is usually used to approximate it

$$\frac{\partial \log P(\mathbf{v})}{\partial W_{ij}} \approx \langle v_i h_j \rangle^0 - \langle v_i h_j \rangle^k \tag{4}$$

where $\langle \cdot \rangle^m$ represents the average value at contrastive divergence iteration m.

To form a DBN, RBMs are stacked on top of each other, where the visible layers of higher RBMs are the hidden states of the previous RBM. In unsupervised feature

learning, the n RBMs are "unrolled" to form a $2n-1$ layer directed encoder-decoder network that can be trained with backpropogation [6]. To ensure dimensionality reduction (and non-trivial feature learning), the DBN acts like a deep autoassociator, with progressively smaller encoding layers.[2]

Our hypothesis in using DBNs for anomaly detection is that the DBN will not see many (or perhaps any at all) anomalous training samples and will thus poorly learn them, evinced by its poor reconstruction of them in a testing set. Roughly, the higher the reconstruction error, the more "unusual" the input sample. This reconstruction error for each sample can be used with a threshold to produce a detector. We use the root-mean-squared error (RMSE), averaged over the D dimensions of the data, to represent DBN reconstruction error of sample $\mathbf{x}$ and the DBN's reconstruction of it, $\mathbf{y}$: $\mathrm{RMSE}(\mathbf{x}) = \sqrt{\frac{1}{\mathrm{D}}\sum_{\mathrm{j}=1}^{\mathrm{D}}(\mathrm{x_j}-\mathrm{y_j})^2}$ Selecting an RMSE threshold k demarcating the anomaly boundary can be viewed as another model parameter, which we search along with DBN parameters using a validation set with labeled anomalies.

C. SVM Anomaly Detection

We compared the DBN anomaly detector to a state-of-the-art anomaly detector, the one-class Support Vector Machine (SVM) [7], which has been used in a variety of applications, including seizure detection in intracranial EEG [8]. Briefly, for N training samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ drawn from $\mathcal{X}$, the hyperplane is found that separates the data from the origin with maximum margin by solving the quadratic programming problem

$$\min_{\mathbf{w},\rho,\xi} \frac{1}{2} \parallel \mathbf{w} \parallel^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho \tag{5}$$

$$\text{subject to} \qquad \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_{\mathrm{i}}) \geq \rho - \xi_{\mathrm{i}} \quad \text{and} \quad \xi_{\mathrm{i}} \geq 0$$

where ϕ maps the input space $\mathcal{X}$ to the feature space F, $\mathbf{w} \in F$ and $\rho \in \mathbb{R}$ are model parameters, $\xi_i \in \mathbb{R}^N$ are the slack variables, and ν is the asymptotic fraction of anomalies, a parameter that can incorporate a priori domain knowledge. The signum decision function

$$f(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) - \rho) \tag{6}$$

labels anomalies with -1 and the rest +1. We use an SVM with the radial basis function kernel, $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2}$ with $\gamma > 0$, given its ability to model many different types of data.

Since detector precision and recall usually improve at the expense of each other, we use the F_1 measure, which incorporates both, to assess the performance of the detectors

$$F_1 = 2\frac{\mathrm{recall} \cdot \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}} \tag{7}$$

[2]But even when the number of hidden units is larger than the numebr of visible units, weight regularization during RBM training usually prevents trivial visible-hidden layer mappings.

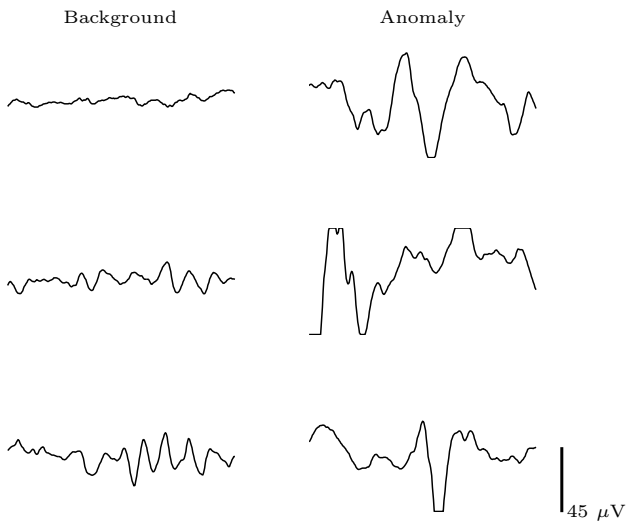

Figure 1. Random examples of background (left) and anomaly (right) 1-second waveforms. Samples are truncated at $\pm$ 45 μV to provide a fixed bounds for scaling to values between 0 and 1, affecting 2.43% of the datapoints.

D. Overview

In Section II, we describe the EEG dataset used in these experiments. In Section III, we compare the training and performance of the DBN and SVM anomaly detectors. In Section IV, we show how the DBN's reconstruction error can be used to visualize anomalous samples in large amounts of EEG. In Section V, we show how top-layer DBN activations from individual anomalous samples can give insights into the anomalous features of those samples.

II. EEG Datasets

We used scalp EEG data sampled at 256 Hz from 11 patients in continuous monitoring at the Hospital of the University of Pennsylvania intensive care unit while receiving therapeutic hypothermia treatment following cardiac arrest-induced coma. All EEG channels were recorded referentially and localized with bipolar montages [9]. We gathered both an unlabeled set used for training and a separate labeled set used for validation and testing. For the unlabeled set, we randomly sampled to create a dataset of 500,000 individual channel-seconds[3].

For the labeled set, we randomly selected 50 2-minute segments from areas in patient records with high-prevalence of unusual activity in order to get enough unusual samples for validation and testing[4]. A clinical neurophysiologist

[3]Average patient contribution to unlabeled dataset: mean and standard deviation $= 9.1\% \pm 11.0\%$, max $= 41.5\%$, min $= 3.4\%$

[4]Patient contribution to labeled dataset: mean and standard deviation $= 9.1\% \pm 3.5\%$, max $= 18.0\%$, min $= 5.9\%$

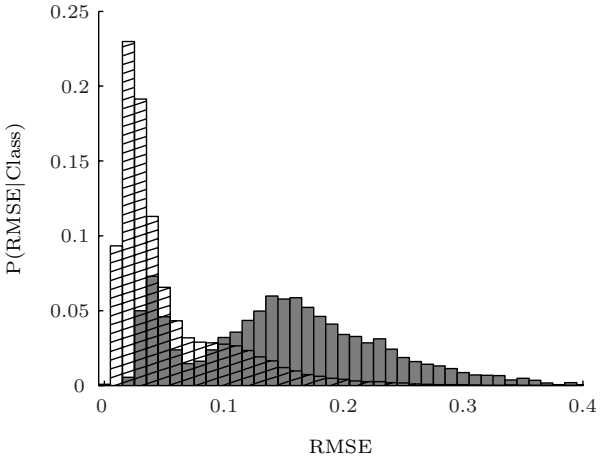

Figure 2. Histogram estimates of the class-conditional probability density functions of the DBN reconstruction error (RMSE) for background (hatched) and anomalous (grey) classes.

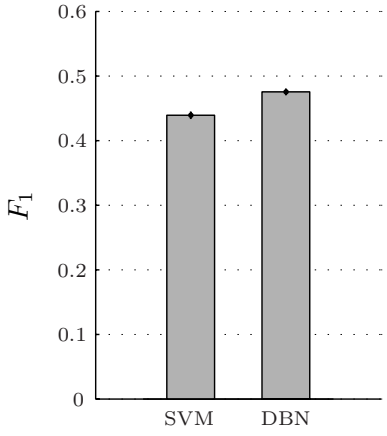

Figure 3. Average F_1 across 10 partitions for the top SVM and DBN anomaly detectors with standard deviation error bars. The SVM detector achieved an F_1 of 0.4390 $\pm$ 0.0049 and the DBN an F_1 of 0.4752 $\pm$ 0.0044.

(R.M.) labeled individual channel-seconds as either background or one of 4 clinically relevant waveform classes[5], often labeling individual channels in the same second differently since the clinically relevant waveforms may only be present in one or a few channels at a given time. The reviewer could see all the channels of the entire 2-minute segment and also had access to his clinical notes for that patient. We lumped the clinically relevant waveforms classes together to make the anomaly class for a total of 8.42% of the samples labeled as anomalies. We exploited the rarity of the relevant waveforms in treating them as anomalous, but in other domains, the "interesting" samples may of course not also be the rare ones. It is important to note that even with this context, the line between background and anomalous can be difficult to determine. We randomly split these labeled samples into 10 partitions of validation and testing sets, each containing 50,900 samples.

As others have done in EEG classification tasks [11], we used the original signal as the input to the detectors. Although DBNs and SVMs can both handle continuous-valued inputs, we introduced a degree of constraint to faciliate learning by linearly scaling the inputs to values between 0 and 1. Data points above and below $\pm$ 45 μV were truncated to 1 and 0, respectively. This truncation affected only 2.43% of data points across all the samples used.

Figure 1 shows random examples of three background and three anomalous signals. The background signals in the left column are consistent with the clinical norm of low-voltage background (e.g.± 15 μV) for these patients. The anomalies show higher voltage (e.g. ± 40 μV). The shape of the waveforms, in addition to the amplitude, also determines if it is anomalous or not, preventing the use of a simple voltage threshold for separating the two classes.

[5]Spike & Sharp Wave, Generalized Periodic Epileptiform Discharge & Triphasic, Periodic Lateralized Epileptiform Discharge, and Eye Blink Artifact [10]

III. Comparing Anomaly Detectors

A. Training

Both DBN and SVM detectors were trained on the same set of 500,000 unlabeled samples. The RBM layers of the DBN were first individually trained using 1-step Contrastive Divergence [6] and then joined together and trained with backpropogation, minimizing the log loss of the reconstruction [3].

We used the labeled validation set of 50,900 samples to do model selection on the DBN and SVM[6]. The optimal DBN size was 250-100-50-5 with RMSE threshold $k = 0.14$, and the top SVM experiment had $\gamma = 1$ and $\nu = 0.04$. One advantage of the DBN method is that determining the best fraction of anomalies (and thus the optimal threshold) requires a very fast search (e.g. 10 seconds) after the main DBN training, whereas determining the best ν for the one-class SVM requires training individual SVMs, usually a much more computationally-intensive task (e.g. 3-8 days per SVM).

Figure 2 shows, for one of the labeled partitions, histogram estimates of the class-conditional probability density functions of root-mean-squared DBN reconstruction error for the background and anomaly classes. In this EEG dataset,

[6]Parameter searches were informed by initial experiments. For the DBN, we searched only layer-size parameters H_ℓ for layer ℓ ($H_1 \in \{250\}$, $H_2 \in \{100, 250\}$, $H_3 \in \{50, 100\}$, $H_4 \in \{2, 5, 15\}$) and RMSE error anomaly threshold ($k \in \{0.01, 0.02, \ldots, 0.49, 0.5\}$). For the SVM, we searched the anomaly fraction ($\nu \in \{0.04, 0.08, 0.12, 0.16\}$) and the size of the radial basis function ($\gamma \in \{0.01, 0.1, 1, 10\}$). We used the LIBSVM package [12] for the SVM experiments.

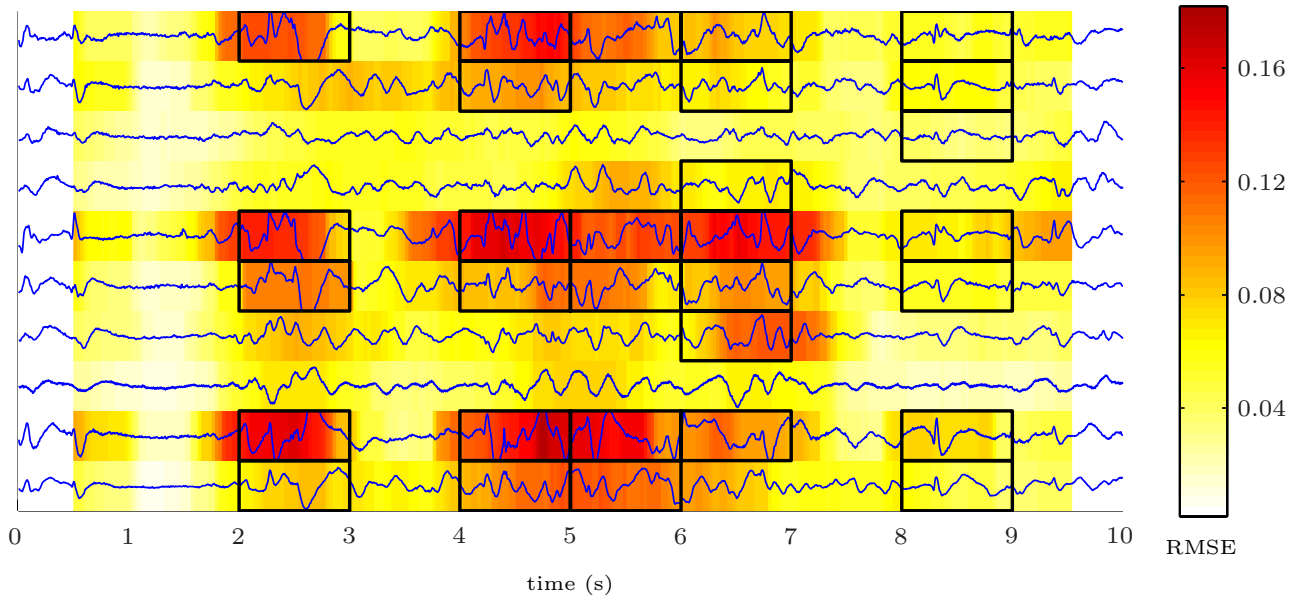

Figure 4. Large-scale anomaly visualization using DBN RMSE on a 10-second clip of 10 EEG channels. The color behind a point on a given channel represents the windowed (1 second, 62.5 ms (8 data points) window overlap) RMSE centered around that point. More anomalous areas of the signal have higher RMSE and are redder in color. Samples which a human reviewer independently labeled as anomalies are boxed.

we notice that the distributions significantly overlap, making it impossible to completely separate the two classes. We also notice that the anomaly distribution is somewhat bimodal, indicating that a subset of the anomalies is fairly-well represented by the DBN. Since some of the anomalous waveforms were of a periodic nature, it is possible that enough were present in the unlabeled data or where similar enough to the background that the DBN learned to represent them reasonably well. This figure also confirms our intuition that DBNs with low average RMSE (like the best performing 250-250-100-15 architecture) are not necessarily best suited to the anomaly detection task when considering a metric like F_1 that balances precision and recall. These more "powerful" DBNs are better at modeling a larger range of input signals, including more of the anomalies, leading to their poorer RMSE separation.

B. Results

Figure 3 shows the F_1 performance of the SVM and DBN anomaly detectors. While the performance of both detectors is similar, the DBN clearly outperforms the SVM by a small margin. The SVM had a higher recall (0.6529 vs. 0.5514) but lower precision (0.3307 vs. 0.4175) than the DBN. We hesitate to draw any large-scale conclusions about the two methods from these results except that the DBN's threshold-based parameter search is much faster than the SVM's ν search, as previously mentioned, making it easier to find an optimal value in a constrained amount of time.

Although both the DBN and one-class SVM both take considerable training time (a few days to over a week for the high ν SVMs on a 2.66 GHz Intel Xeon Dual-Core CPU with 12 GB RAM), they are not equally affected by the a priori distribution of anomalies in the training data. Higher fractions of anomalies will only affect the DBN RMSE threshold search, which has negligible computational cost, whereas they influence optimal values of ν for the SVM, which can drastically affect training time as the number of training samples increases [7]. The training time of a DBN is thus much less influenced by the underlying distribution of training data than that of an SVM.

The DBN query time—the time to classify samples as background or anomaly—is also orders of magnitude faster than that of the the SVMs we tested, which makes sense since it requires only simple matrix multiplication. While a formal handling of this observation is beyond the scope of this paper, in our experiments, the DBNs generally took around 10 seconds to classify the roughly 50,000 samples in a validation set, whereas the SVMs took many hours to do the same. This observation suggests that DBNs may be better-suited for real-time detection applications involving large quantities of data, such as multichannel EEG detection tasks. However, a thorough analysis of computational costs of the various implementations of the two methods remains to been completed.

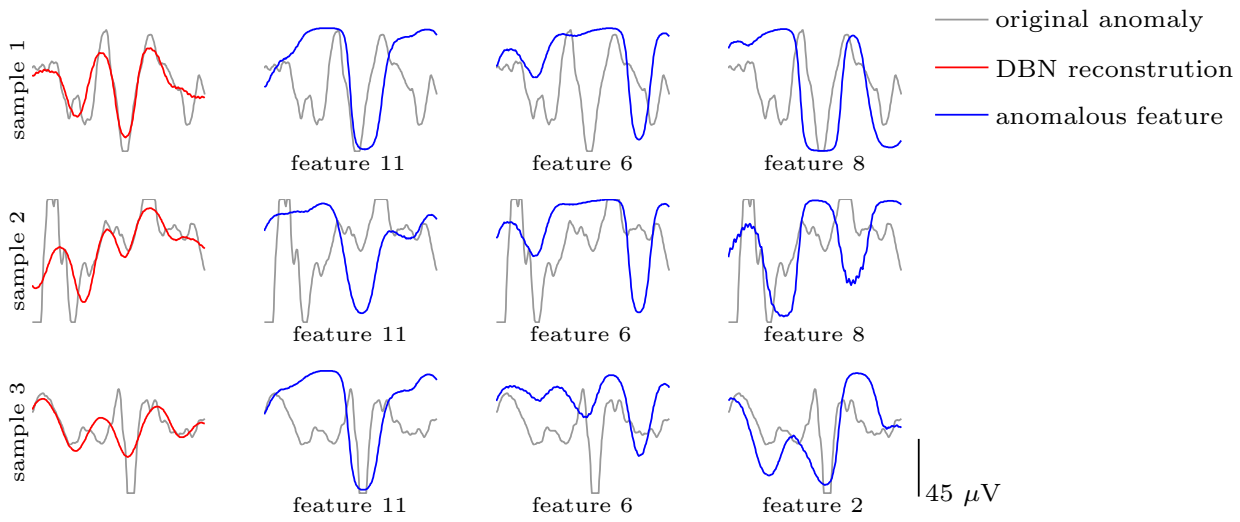

Figure 5. The DBN reconstructions (red) and top three anomalous features (blue, in order from left to right) for three random anomalies (gray,rows). Each anomalous feature is produced by propagating the difference between the mean of a top-level unit and its value for that sample down through the reconstruction side of the DBN.

IV. Large-Scale Anomaly Visualization

The continuous-valued RMSE of the DBN can be used to characterize the collective degree of anomaly in a large dataset. In clinical EEG review, for example, physicians almost always look at many EEG channels simultaneously (often 10 second screens). In Figure 4, we demonstrate how such a view combined with the DBN RMSE of individual parts can yield an overall picture of how anomalous a block of data is.

The figure shows 10 channels of a 10-second block of EEG from the validation set. For each channel, we first encoded and reconstructed successive seconds using the 250-100-50-5 DBN, yielding reconstruction samples, which we concatenated to form a 10-second long reconstruction signal corresponding to the 10-second long original signal. We then computed the sliding RMSE between the original and reconstruction signals using a symmetric 1-second sliding window with 62.5 ms (8 data points) overlap).

To better illustrate how the RMSE correlates with the human anomaly labels, we have boxed the samples in Figure 4 that the human reviewer independently labeled as anomalies. The areas with high-anomaly content (orange-red) generally correspond to samples marked as such by the human. We note, however, that the RMSE is low in low-voltage, fine-detail anomalies like those on the right of the figure. The DBN seems to have "learned" these waveforms adequately enough to be able to reconstruct them. Without the human labels, the DBN RMSE provides a very informative measure by which to judge anomalous behavior.

Such large-scale visualizations of large numbers of individual samples allow a human to quickly understand what the DBN thinks is anomalous and what is not. Proper detection of anomalous waveforms often requires a highly-trained neurophysiologist. Improved visualization methods like the one we present here would allow more rapid and accurate detection by both trained neurophysiologists and others (e.g. nurses, technicians) in the patient's health care team not trained in neurophysiology.

V. Observing Anomalous Features

Visualizations such as Figure 4 prompted us to explore what unusual features anomalies contain. DBNs offer a natural way of observing sample features via the DBN's hierarchical layers, which can also be thought as hierarchical feature detectors. Thus, the top-level units can be thought of as encoding the highest-level features. The auto-encoding nature of DBNs allows us to visualize what the activations of those top-level units looks like in the original sample space.

To visualize features from a specific anomalous sample, we use the following procedure:

1) Estimate the distribution of the activations of the top-level DBN units over the entire training set using a histogram.
2) Calculate the top-level activation values for a specific anomalous sample. Calculate the difference between these values and the mean distribution values for each unit and normalize this difference by the standard deviation of the unit's distribution, yielding the z-score.
3) Sort the Xactivation differences in descending order.
4) For the first n differences

a) Zero all the top-level activation values except that of the nth unit, which takes the z-score activation difference of that unit.
b) Propagate these all-zero-but-one activations down through the reconstruction part of the DBN to visualize the z-score activation difference for the specific nth feature unit.

We follow these steps in Figure 5 to produce the top three anomalous features (in order from left to right) for the three anomalies shown in Figure 1. We show the DBN reconstruction (red) as well as the anomalous features (blue) overlaid on the original anomaly signal (gray) for comparison. The top-level units used for each sample are not necessarily the same (or in the same order) although in this figure they generally are. The anomalous features produced seem to have some correlation with aspects of the original waveform, indicating that they connote some relevant information about the original signal. Furthermore, the anomalous features in the first and second columns (from the left) bear some resemblance to the others in each column, even though the original waveforms look somewhat different. This observation suggests that these features, with large-amplitude toughs, are indicative of anomalous samples.

In generating such figures, it is necessary to use a DBN that models the data reasonably well to produce meaningful features. The same DBN used in the original anomaly detection task most likely is not appropriate because it only models the data moderately well (so that it can produce high RMSE for anomalies). In Figure 5, we use a 250-250-100-15 DBN, which had the lowest average validation RMSE of all the DBNs trained, but it might be more appropriate to train a much larger and more powerful DBN for the sole purpose of visualizing anomalous features.

Observing the most anomalous features of individual and possibly groups of samples, especially in conjunction with the type of gross visualization we describe in Section IV, gives physicians and researchers an unbiased method by which to examine unusual and potentially novel (with respect to clinical knowledge) features of EEG. We believe that these types of visualization and feature discovery have merit in many other application domains where the quantity and dimensionality of the data is high, such as image processing, geosensing, and industrial manufacturing.

VI. Conclusion

We show that Deep Belief Nets can be used for semi-supervised anomaly detection. In our clinical EEG application, DBNs outperform the state-of-the-art, a one-class Support Vector Machine, using F_1 as the performance metric. We note that DBNs have several other advantages that make them attractive to a wide array of anomaly detection domains: training scales efficiently to a large training set, the fraction of anomalies in the data has a negligible effect on the training time (unlike the SVM), and they have fast query time. We show that the DBN's reconstruction RMSE can measure the degree of anomaly, making large-scale anomaly visualization straightforward. Finally, we explore how the hierarchical feature-layers of a DBN can be used to extract the anomalous features of individual samples. We believe that this work demonstrates not only that DBNs can be used as high-performance anomaly detectors but also that they can participate in shaping our understanding of the very anomalies they detect.

Acknowledgment

This work is supported by NIH grants 5T90DA022763-04 and NS061779-02.

References

[1] K. Muller, M. Krauledat, G. Dornhege, G. Curio, and B. Blankertz, "Machine learning techniques for brain-computer interfaces," *Biomed Tech*, vol. 49, no. 1, pp. 11–22, 2004.

[2] V. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, Oct. 2004.

[3] G. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 504, pp. 504–507, 2006.

[4] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[5] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards AI," *Large-Scale Kernel Machines*, no. 1, pp. 1–41, 2007.

[6] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[7] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution." *Neural computation*, vol. 13, no. 7, pp. 1443–71, Jul. 2001.

[8] A. Gardner, A. Krieger, G. Vachtsevanos, and B. Litt, "One-class novelty detection for seizure analysis from intracranial EEG," *J Machine Learning Research*, vol. 7, p. 1044, 2006.

[9] J. Ebersole and T. Pedley, *Current practice of clinical electroencephalography*, 3rd ed. Lippincott Williams & Wilkins, 2003.

[10] D. J. Chong and L. J. Hirsch, "Which EEG patterns warrant treatment in the critically ill? Reviewing the evidence for treatment of periodic epileptiform discharges and related patterns." *J Clin Neurophys*, vol. 22, no. 2, pp. 79–91, Apr. 2005.

[11] B. Blankertz, G. Curio, and K. Müller, "Classifying single trial EEG: Towards brain computer interfacing," *NIPS 2002*, no. c, 2002.

[12] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," 2001.