




Unsupervised learning: application to epilepsy

Aprendizaje no supervisado: aplicación en epilepsia

Gabriel Mauricio Martínez-Toro¹ , Dewar Rico-Bautista² , Efrén Romero-Riaño³ 
Paola Andrea Romero-Riaño⁴

¹Doctorado en Ingeniería, Universidad Autónoma de Bucaramanga, Bucaramanga, Colombia.

²Programa de Ingeniería de Sistemas, Universidad Francisco de Paula Santander Ocaña, Ocaña, Colombia.

³Grupo de Investigación en Gestión de la Innovación Tecnológica y del Conocimiento (INNOTEC), Universidad Industrial de Santander, Bucaramanga, Colombia.

⁴Doctorado en Salud Pública con mención en Sistemas y Servicios de Salud, Universidad de Ciencias Empresariales y Sociales UCES, Buenos Aires, Argentina

gmartinez714@unab.edu.co, dwricob@ufps.edu.co, efren.romero@saber.uis.edu.co, pagla220@hotmail.com

(Recibido: 24 enero 2019; aceptado: 17 julio 2019)

Abstract

Epilepsy is a neurological disorder characterized by recurrent seizures. The primary objective is to present an analysis of the results shown in the training data simulation charts. Data were collected by means of the 10-20 system. The “10-20” system is an internationally recognized method to describe and apply the location of scalp electrodes in the context of an EEG exam. It shows the differences obtained between the tests generated and the anomalies of the test data based on training data. Finally, the results are interpreted and the efficacy of the procedure is discussed.

Keywords: Epilepsy; Deep learning; Automatic learning; Auto-encoding.

Resumen

La epilepsia es uno de los trastornos neurológicos comunes caracterizado por convulsiones recurrentes. El objetivo principal de este artículo es dar a conocer el análisis de los resultados presentados en las gráficas de simulación de los datos de entrenamiento. Los datos fueron recolectados mediante el sistema 10-20. El sistema “10-20” es un método reconocido internacionalmente, este describe la ubicación de electrodos en la cabeza para una prueba de EEG. Se muestran las diferencias obtenidas entre las pruebas generadas con las anomalías de los datos de prueba a partir de los datos de entrenamiento. Finalmente, se interpretan los resultados y se discute sobre la eficacia del procedimiento.

Palabras clave: Epilepsia, Aprendizaje profundo, Aprendizaje automático, Auto codificación.

1. Introduction

Epilepsy is a neurological disorder characterized by recurrent seizures (Beatriz Pérez Salazar & Lillia Hernández López, 2007). These seizures are seen as a sudden abnormal function of the body, often with a loss of consciousness, increased muscle mass activity or abnormal sensation (Kuremoto, Kimura, Kobayashi, & Obayashi, 2014). Epilepsy is characterized by recurrent seizures in which electrical anomalies in the brain's activity cause an altered perception or behavior (López-Meraz et al., 2009). Patients experience various symptoms during seizures, depending on the location and extension of the affected brain region (P. Mirowski, Madhavan, LeCun, & Kuzniecky, 2009).

Generalized seizures involve almost the whole brain, while partial seizures originate in a specific region of the brain and remain restricted to that region (Cruces, 2014). Epileptic seizures may cause negative



Cite this work as Martínez-Toro G., Rico-Bautista D., Romero-Riaño E., Romero-Riaño P. (2019). Unsupervised learning: application to epilepsy. *Revista Colombiana de Computación*, 20(2), 20-27. <https://doi.org/10.29375/25392115.3718>

physical, psychological and social consequences, including loss of consciousness, injuries and sudden death (Aarabi & He, 2012). So far, the specific cause of epilepsy in individuals is not known, and the mechanisms behind seizures are poorly understood (Fuertes, López, & Gil, 2007). Therefore, efforts to diagnose and treat the disorder are extremely important (P. W. Mirowski, Lecun, Madhavan, & Kuzniecky, 2008).

The EEG signal is one of the most used in bioinformatics because of its wealth of information on human activity. The EEG has been an important clinical tool to evaluate human brain activity (P. W. Mirowski, Madhavan, & Lecun, 2007). EEG monitoring provides valuable information about candidates that suffer from epilepsy (Valencia et al., 2016). EEG recording of patients suffering from epilepsy shows two categories of abnormal activity: abnormal interictal signal recorded between epileptic attacks; and ictal, the activity recorded during an epileptic attack (Aliper et al., 2016). Clinical EEG of the scalp is used to diagnose and guide therapy for a variety of neurological disorders that include acute attacks and cerebral ischemia after a stroke and cardiac arrest (Soleimani-B., Lucas, N. Araabi, & Schwabe, 2012).

Clinical EEG monitoring often employs automatic algorithms to detect epileptiform discharges and activity resembling an attack, but most of these tools are plagued by low performance and high false positive rates that limit its clinical usefulness (Wang & Shang, 2014). To that end, we present an approximation based on the use of autoencoder techniques. The article is divided in three parts: the first presents the materials and methods, the second states the results of the simulations, with a small discussion, and the last asserts the conclusions.

2. Materials and methods

2.1 Machine Learning

There are various conceptual frameworks for Machine Learning. Here we summarize some aspects related to the architectures used. Following is a brief description of machine learning algorithms:

Logistic regression is a well-established classification technique that is widely used in epidemiological studies. It is generally used as reference in comparison with other medical data analysis techniques. Logistic regression is also known as logit regression, maximum entropy model (MaxEnt), or log-linear classifier (Dreiseitl & Ohno-Machado, 2002).

Linear discriminant analysis is a linear classifier. These classifiers are attractive because they have closed form solutions which can be easily calculated. They can be used to do supervised dimensionality reduction, projecting entry data on a linear subspace consisting of directions that maximize the separation between classes (Dudoit, Fridlyand, & Speed, 2002).

K-nearest neighbors algorithm (KNN) is a classic learning algorithm based on instances in which a new case is classified based on the known class of the nearest neighbor by means of a majority of samples. The principle behind the KNN methods is to find a pre-defined number of training samples closest in distance to the new point, and to predict their label. The number of samples can be a user-defined constant (k-nearest neighbor learning) or vary depending on the local density of points (radius based nearest neighbor learning). This type of learning is also called lazy learning because there is no step to construct models and all the procedure information (in other words, the search for the nearest neighbor) is done directly during the prediction (Dreiseitl & Ohno-Machado, 2002).

The *Naive Bayes* methods are a set of supervised learning algorithms based on the application of Bayes' theorem with the supposition of interdependence between each pair of characteristics. The algorithm works well with heterogeneous data types and also with lost values due to the independent treatment of each predictive variable in the construction of the model (Griffis, Allendorfer, & Szaflarski, 2016).

Support-vector machines are a set of supervised learning methods used to detect classification, regression and others (Chisci et al., 2010). SVM's main advantages include: effective in high dimension spaces, still effective in cases where the number of dimensions is higher than the number of samples, and efficient from the point of view of memory (Chang & Lin, 2011).

The multilayer backpropagation perceptron is the main artificial material of the neural network. When there is no hidden layer in the network, this algorithm is equivalent to logistic regression, but it can solve more difficult problems with a more complex architecture network. The price of using more complex architectures is that it produces models that are more difficult to interpret. In addition, it can be computationally more expensive.

Random forest is a machine learning algorithm of the ensemble family of algorithms methods that creates multiple models (called weak learners) and combines them to make a decision in order to increase the accuracy of the prediction. The main idea here is to create a “forest” of random decision “trees” and use it to classify a new case. Each tree is generated using a random tree of the candidate’s predictor variables and a sub-set of random variables. This algorithm can also be used to estimate variable relevance.

The *k-means* algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires that the number of clusters be specified. It adapts well to a large number of samples (Tsai, 2014).

Decision trees are a non-parametric supervised learning method used for both classification and regression tasks (Langkvist, Karlsson, & Loutfi, 2014). The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features (Kurzynski, Krysmann, Trajdos, & Wolczowski, 2016).

2.2 “10-20” System

Data are collected by means of the 10-20 system. The “10-20” system is an internationally recognized method to describe and apply the location of scalp electrodes in the context of an EEG exam (Alshebeili, Alshawi, Ahmad, & El-samie, 2014; Escalona-Morán, Cosenza, Guillén, & Coutin, 2007). The system is based on the relationship between the location of an electrode and the underlying area of cerebral cortex. The numeric term “10” and “20” means the distances between adjacent electrodes. There is either 10% or 20% of the total front-back or right-left distance of the skull (Garg & Narvey, 2013). Figure 1 illustrates the 10-20 system used.

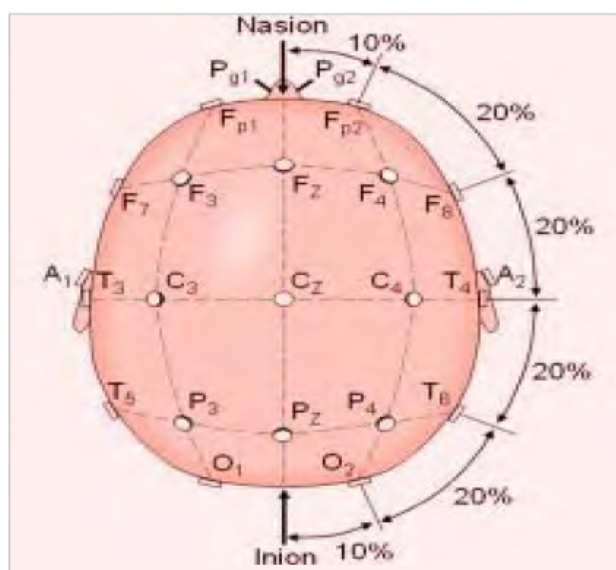


Figure 1. Location of electrodes for an EEG test (Garg & Narvey, 2013)

This analysis is based on a procedure that allows to characterize the quantity of synchronization and clustering that occurs in coupled chaotic oscillators subject to common noise, and it applies these concepts to the EEG signals of healthy subjects and epileptic patients (Escalona-Morán et al., 2007). A system of coupled oscillators is a set of individual oscillators that are interconnected. The coupled oscillators model may be applied to both mechanical systems and solid atomic models (Garg&Narvey, 2013). Just as each oscillatory system has an associated oscillation frequency characteristic, a system with multiple coupled oscillators has a set of oscillation modes with defined frequency characteristics. This property is used to identify the characteristics (Escalona-Morán et al., 2007).

Autoencoders are used to process the signals obtained. Autoencoders for EEG input signals produce a reconstructed signal of the nearest possible input signal, giving a fixed number of layers. In this case, they

use two neurons each with 50 layers and results are generated through iteration in 100 epochs (D. Wulsin, Blanco, Mani, & Litt, 2010).

Our hypothesis is that through autoencoders, the machine learns signal types that are more frequent in training data, producing better reconstructions thereof. Similarly, “unusual” or anomalous signals will rarely occur in training data, preventing the generative graph models from learning and also reconstructing them. Although some aspects of the most common signals appear to be harder to learn by autocoding (i.e., components with higher amplitude and lower frequency), it was identified that through this system the majority of the aspects of the common signals are learned better than the less common signals (D. F. Wulsin, Gupta, Mani, Blanco, & Litt, 2011). Additionally, it estimates the precision with which the proposed scheme transforms an x input sample into a z reconstruction through autocoding.

2.3 Input data

Input data correspond to data from the EEG that present a set of 10 columns that represent readings from the following points: Occipital (O), Frontal (F), Parietal (P), depending on the 10-20 system focus. The set of data are comprised of a total of around eight thousand data and is based on shared public information (Escalona-Morán et al., 2007). This set has four subsets.

The research design designates as an input set or training data a total of 80% of the EEG data, and the remaining 20% is taken as validation data. A simulation of training data is executed through the use of the C4 and O1 channels. A comparison of the results of the simulation of the training data, the prediction of anomalies and prediction chart are presented in the results section.

3. Results and discussion

The transformation of the input data in the set of characteristics is called extraction of characteristics. If the extracted characteristics are carefully chosen, it is expected that from these we will extract relevant information from the input data to conduct the desired characterization (Garg & Narvey, 2013). These are extracted to help distinguish between normal and epileptic signal (P. Mirowski et al., 2009; P. W. Mirowski et al., 2008, 2007). Following are the charts of the training data simulation. It is possible to contrast the result obtained between the C4 (see Figure 2) and O1 (see Figure 3) group of data

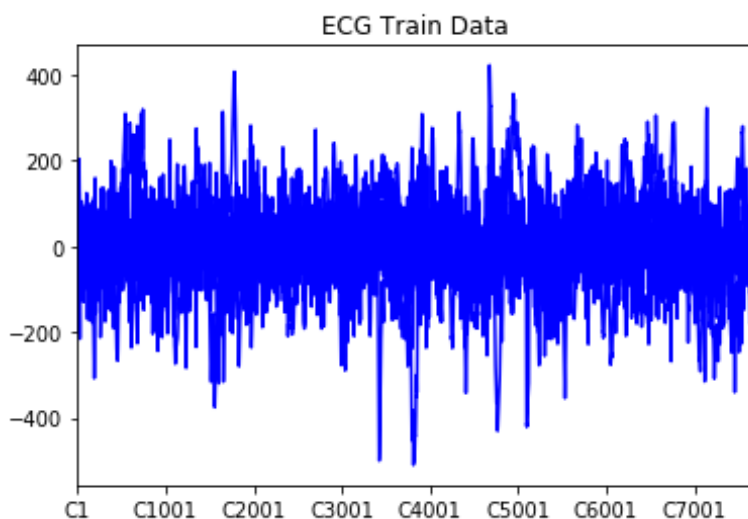


Figure 2. Results of the simulation: C4 training data.

Source: Prepared by the authors using the Anaconda Spyder

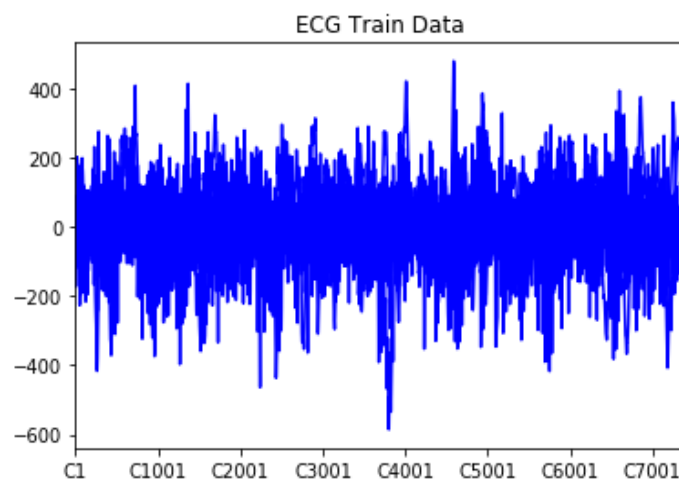


Figure 3. Results of the simulation: O1 training data.
Source: Prepared by the authors using the Anaconda Spyder

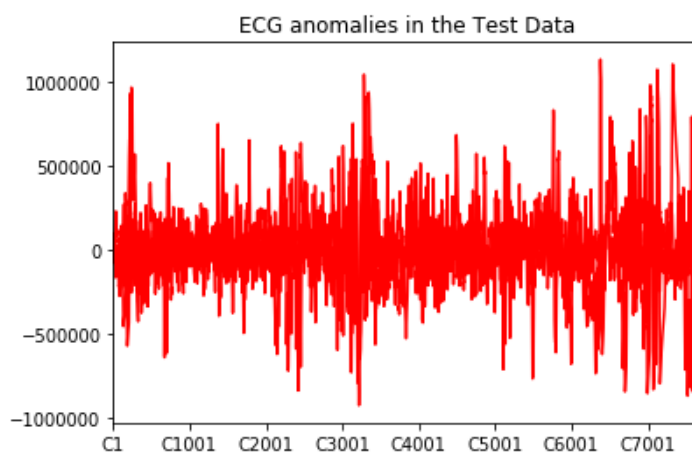


Figure 4. Anomalies in the reconstruction of C4 test data
Source: Prepared by the authors using the Anaconda Spyder

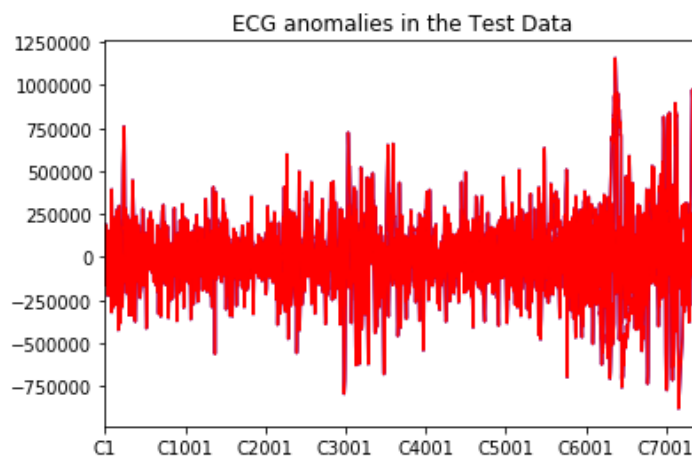


Figure 5. Anomalies in the reconstruction of O1 test data
Source: Prepared by the authors using the Anaconda Spyder

Figures 4 and 5 show the differences obtained between the tests generated with the anomalies in the test data. Table 1 is presented as a measure to interpret the results, which displays the data reconstruction errors from groups C4 and O1.

Table 1. Errors in the reconstruction of C4

Reconstruction.MSE Rank			Reconstruction.MSE Rank		
6	4.363190e+06	1.0	6	1.438027e+06	1.0
5	1.979819e+06	2.0	5	1.116391e+06	2.0
7	1.037147e+06	3.0	7	1.029054e+06	3.0
1	8.369630e-06	4.0	8	4.047404e+05	4.0
2	7.464590e-06	5.0	1	9.567050e-06	5.0
4	3.834310e-06	6.0	4	8.281564e-06	6.0
3	2.652784e-06	7.0	2	3.660314e-06	7.0
0	1.585078e-06	8.0	3	3.337196e-06	8.0
Reconstruction.MSE Rank			Reconstruction.MSE Rank		
6	4.363190e+06	1.0	6	1.438027e+06	1.0
5	1.979819e+06	2.0	5	1.116391e+06	2.0
7	1.037147e+06	3.0	7	1.029054e+06	3.0

Source: Prepared by the authors using the Anaconda Spyder

The EEG signal is one of the most used signals because of its wealth of information on human activity. The EEG has been an important clinical tool in evaluating human brain activity. EEG monitoring provides valuable information about candidates that suffer from epilepsy. The electroencephalogram record of a patient that suffers from epilepsy shows two categories of abnormal activity: interictal, an abnormal signal recorded between epileptic seizures; and ictal, the activity recorded during an epileptic crisis.

The training was conducted by progressively refining the hypothesis function. We kept a record of events occurred and compared them with historical data to try to detect anomalies. At the end of the process (see Figures 6 and 7), the epileptic and normal cases were diagnosed.

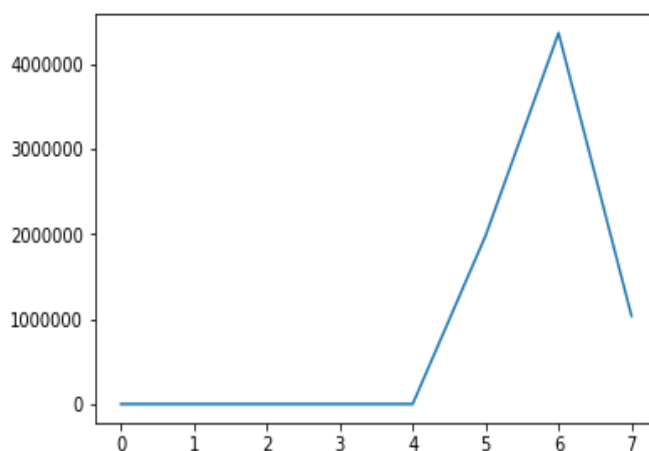


Figure 6. C4 set prediction chart

Source: Prepared by the authors using the Anaconda Spyder

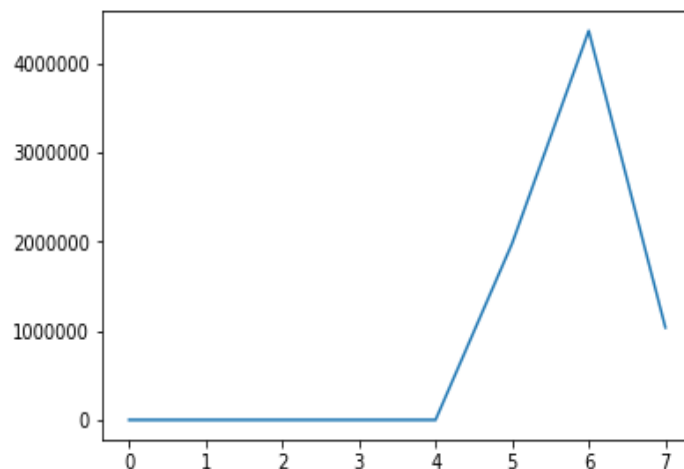


Figure 7. O1 set prediction chart

Source: Prepared by the authors using the Anaconda Spyder

4. Conclusions

This article illustrates the application of a machine learning technique to detect the epileptic disorder, which is efficient to distinguish normal and epileptic signals. The autoencoder generated may be used naturally in measuring anomalies in EEG signals.

A comparison of raw, unprocessed data with the randomly selected characteristics showed that raw data produce a comparable classification and better yield of the anomalies measurement.

We observed that the anomalies showed frequencies in the order of 1000000 in contrast with the range of 400 in C4; similarly, in O1 the anomalies in the reconstruction of the data showed frequencies of 1250000 in contrast with the range of 400 of the training data.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

References

- Aarabi, A., & He, B. (2012). A rule-based seizure prediction method for focal neocortical epilepsy. *Clinical Neurophysiology*, 123(6), 1111–1122. <https://doi.org/10.1016/j.clinph.2012.01.014>
- Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., Zhavoronkov, A., & Albuquerque, N. (2016). HHS Public Access, 13(7), 2524–2530. <https://doi.org/10.1021/acs.molpharmaceut.6b00248>.Deep
- Alshebeili, S. A., Alshawi, T., Ahmad, I., & El-samie, F. E. A. (2014). EEG seizure detection and prediction algorithms : a survey. *EURASIP Journal on Advances in Signal Processing*, 183(1), 1,21. <https://doi.org/10.1186/1687-6180-2014-183>
- Beatriz Pérez Salazar, Á., & Lillia Hernández López, D. (2007). Epilepsia: aspectos básicos para la práctica psiquiátrica Epilepsia: aspectos básicos para la práctica psiquiátrica Title: Epilepsy: Basic Aspects for the Practice of Psychiatry. *Rev. Colomb. Psiquiat*, XXXVI XXXV(1), 175–186.
- Chang, C.-C., & Lin, C.-J. (2011). Libsvm. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1–27. <https://doi.org/10.1145/1961189.1961199>
- Chisci, L., Mavino, A., Perferi, G., Sciandrone, M., Anile, C., Colicchio, G., & Fuggetta, F. (2010). Real-Time Epileptic Seizure Prediction Using AR Models and Support Vector Machines. *IEEE Transactions on Biomedical Engineering*, 57(5), 1124–1132. <https://doi.org/10.1109/TBME.2009.2038990>
- Cruces, H. De. (2014). Tipos de crisis epilépticas y pseudocrisis Diferencial characteristics of epileptic seizure and pseudoseizures, 105–107.

- Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35(5–6), 352–359. [https://doi.org/10.1016/S1532-0464\(03\)00034-0](https://doi.org/10.1016/S1532-0464(03)00034-0)
- Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457), 77–86. <https://doi.org/10.1198/016214502753479248>
- Escalona-Morán, M., Cosenza, M. G., Guillén, P., & Coutin, P. (2007). Synchronization and clustering in electroencephalographic signals. *Chaos, Solitons and Fractals*, 31(4), 820–825. <https://doi.org/10.1016/j.chaos.2005.10.049>
- Fuertes, B., López, R., & Gil, P. (2007). Epilepsia. *Tratado de Geriatria Para Residentes*, 519–530.
- Garg, S., & Narvey, R. (2013). Denoising & feature extraction of eeg signal using wavelet transform. *International Journal of Engineering Science and Technology*, 5(06), 1249–1253.
- Griffis, J. C., Allendorfer, J. B., & Szaflarski, J. P. (2016). Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. *Journal of Neuroscience Methods*, 257, 97–108. <https://doi.org/10.1016/j.jneumeth.2015.09.019>
- Kuremoto, T., Kimura, S., Kobayashi, K., & Obayashi, M. (2014). Time series forecasting using a deep belief network with restricted Boltzmann machines. *Neurocomputing*, 137, 47–56. <https://doi.org/10.1016/j.neucom.2013.03.047>
- Kurzynski, M., Krysmann, M., Trajdos, P., & Wolczowski, A. (2016). Multiclassifier system with hybrid learning applied to the control of bioprosthetic hand. *Computers in Biology and Medicine*, 69, 286–297. <https://doi.org/10.1016/j.combiomed.2015.04.023>
- Langkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42(1), 11–24. <https://doi.org/10.1016/j.patrec.2014.01.008>
- López-meraz, M. L., Rocha, L., Miquel, M., Hernández, M. E., Cárdenas, R. T., Coria-ávila, G. A., ... Manzo, J. (2009). Conceptos básicos de la epilepsia. *Revista Médica de La Universidad Veracruzana*, 9(2), 31–37.
- Mirowski, P., Madhavan, D., LeCun, Y., & Kuzniecky, R. (2009). Classification of patterns of EEG synchronization for seizure prediction. *Clinical Neurophysiology*, 120(11), 1927–1940. <https://doi.org/10.1016/j.clinph.2009.09.002>
- Mirowski, P. W., Lecun, Y., Madhavan, D., & Kuzniecky, R. (2008). Comparing SVM and Convolutional Networks for Epileptic Seizure.
- Mirowski, P. W., Madhavan, D., & Lecun, Y. (2007). Time-delay neural networks and independent component analysis for eeg-based prediction of epileptic seizures propagation. *Advancement of Artificial Intelligence Conference*, 1892–1893.
- Soleimani-B., H., Lucas, C., N. Araabi, B., & Schwabe, L. (2012). Adaptive prediction of epileptic seizures from intracranial recordings. *Biomedical Signal Processing and Control*, 7(5), 456–464. <https://doi.org/10.1016/j.bspc.2011.11.007>
- Tsai, C. F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. *ACM Transactions on Intelligent Systems and Technology*, 16(1), 46–58. <https://doi.org/10.1016/j.inffus.2011.12.001>
- Valencia, J. F., Melia, U. S. P., Vallverdú, M., Borrat, X., Jospin, M., Jensen, E. W., ... Caminal, P. (2016). Assessment of nociceptive responsiveness levels during sedation-analgesia by entropy analysis of EEG. *Entropy*, 18(3). <https://doi.org/10.3390/e18030103>
- Wang, D., & Shang, Y. (2014). Modeling Physiological Data with Deep Belief Networks. *International Journal of Education Technology*, 3(5), 505–511. <https://doi.org/10.7763/IJET.2013.V3.326.Modeling>
- Wulsin, D., Blanco, J., Mani, R., & Litt, B. (2010). Semi-supervised anomaly detection for EEG waveforms using deep belief nets. *Proceedings - 9th International Conference on Machine Learning and Applications, ICMLA 2010*, (April 2016), 436–441. <https://doi.org/10.1109/ICMLA.2010.71>
- Wulsin, D. F., Gupta, J. R., Mani, R., Blanco, J. A., & Litt, B. (2011). Modeling electroencephalography waveforms with semi-supervised deep belief nets: Fast classification and anomaly measurement. *Journal of Neural Engineering*, 8(3). <https://doi.org/10.1088/1741-2560/8/3/036015>