Automated Identification of Abnormal Adult EEGs

S. López, G. Suarez, D. Jungreis, I. Obeid and J. Picone
Neural Engineering Data Consortium, Temple University
Philadelphia, Pennsylvania, USA
{silvia.lopez, gabriella.suarez, david.jungreis, obeid, picone}@temple.edu

Abstract— **The interpretation of electroencephalograms (EEGs) is a process that is still dependent on the subjective analysis of the examiners. Though interrater agreement on critical events such as seizures is high, it is much lower on subtler events (e.g., when there are benign variants). The process used by an expert to interpret an EEG is quite subjective and hard to replicate by machine. The performance of machine learning technology is far from human performance. We have been developing an interpretation system, AutoEEG, with a goal of exceeding human performance on this task. In this work, we are focusing on one of the early decisions made in this process – whether an EEG is normal or abnormal. We explore two baseline classification algorithms: k-Nearest Neighbor (kNN) and Random Forest Ensemble Learning (RF). A subset of the TUH EEG Corpus was used to evaluate performance. Principal Components Analysis (PCA) was used to reduce the dimensionality of the data. kNN achieved a 41.8% detection error rate while RF achieved an error rate of 31.7%. These error rates are significantly lower than those obtained by random guessing based on priors (49.5%). The majority of the errors were related to misclassification of normal EEGs.**

I. Introduction

Electroencephalography (EEG), or the recording of the electrical activity of the brain, has become a relatively inexpensive and practical way to demonstrate the physiological manifestations related to conditions such as epilepsy, seizures, sleep disorders and several types of mental status alterations [1]. While the equipment for acquiring EEG data is relatively inexpensive and easy to use, it takes several years of training for a physician to achieve board certification for reading and reporting EEG studies. Many smaller hospitals and emergency medical services may not have a trained neurologist on site. Even larger facilities find it impractical to have certified staff on-site 24/7 for EEG monitoring. Furthermore, longer-term monitoring studies (LTMs) of neurological activity are becoming increasingly important. Each long-term or continuous EEG monitoring study requires a neurologist to review up to 72 hours worth of data, creating a bottleneck for accurate analysis.

The interpretation of an EEG depends heavily on the subjective judgment of the examiner, a situation that could lead to misdiagnosis or missed events in the record [2]. Maintaining an acceptable level of interrater agreement plays a key role in the assessment of the validity of this diagnostic technique. This affirmation is reinforced by the sensitivity levels of the EEG for the diagnosis of conditions such as epilepsy. Essentially, only 50% of the patients with epilepsy show interictal epileptiform discharges (IED) in their first EEG, a number that is reduced in significance by the fact that at least 30% of non-epileptic patients with other conditions or injuries show this behavior in their recordings [3]. Hence, a majority of the patients that present symptoms that could be related to an epileptic disorder must be subject to more than one EEG prior to a diagnosis.

In this sense, the automated classification of an EEG record as normal or abnormal represents a significant step for the reduction of the visual bias intrinsic to the subjectivity of the record's interpretation. Additionally, the assisted interpretation of the background patterns existing in the signal could help save neurologists time in their daily routine, easing some of the service pressures that arise from increasing demand [3].

The main characteristics of an adult normal EEG are [4]:

(1) Reactivity: Response to certain physiological changes or provocations.

(2) Alpha Rhythm: Waves originated in the occipital lobe (predominantly), between 8-13 Hz and 15 to 45 µV.

(3) *Mu Rhythm:* Central rhythm of alpha activity commonly between 8-10 Hz visible in 17% to 19% of adults.

(4) *Beta Activity:* Activities in the frequency bands of 18-25 Hz, 14-16 Hz and 35-40 Hz.

(5) *Theta Activity:* Traces of 5-7 Hz activity present in the frontal or frontocentral regions of the brain.

Neurologists follow procedures similar to the one summarized in Figure 1 and can usually make this determination by examining the first few minutes of a recording. Hence, in this baseline study, we will focus on examining the first 60 secs of an EEG to calibrate the difficulty of the task.

The visual analysis of an EEG begins with the observation of the occipital alpha rhythm. A decision about the normality of the record heavily depends on the frequency, presence or distortion of this feature [4]. In this sense, the posterior dominant rhythm (PDR) or alpha rhythm that emerges in the posterior regions when the patient's eyes are closed is the main

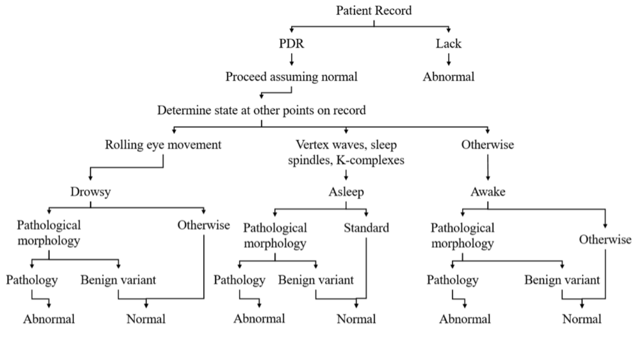

Figure 1. The general process for identifying an abnormal EEG depends heavily on the observation of the PDR.

decisive feature and suggests that detection of this event in an occipital channel of a normal EEG can play a major role in normal/abnormal classification.

An EEG can be considered abnormal for a number of reasons. The most obvious reason, of course, would be the existence of clearly pathological events such as long periods of spike and wave activity, Periodic Lateralized Epileptiform discharges (PLEDs), or Generalized Periodic Epileptiform discharges (GPEDs). The presence of spikes, however, does not guarantee an abnormal EEG. A benign variant is defined as an EEG pattern that is morphologically epileptiform but is not associated with a disease such as epilepsy [3]. Spikes presented in the form of small sharp spikes are considered a benign variant.

II. Experimental Design

In this study we have focused on the TUH EEG Corpus [5] for evaluation. TUH EEG is the world's largest publicly available database of clinical EEG data, comprising more than 28,000 EEG records and over 15,000 patients. It represents the collective output from Temple University Hospital's Department of Neurology since 2002 and is an ongoing data collection project. Approximately 75% of the data represent abnormal EEGs. We selected a demographically balanced subset of the data through manual review that consisted of 202 normal EEGs and 200 abnormal EEGs. These sets were further partitioned into a training set (102 normal/100 abnormal), development test set (50 normal/50 abnormal) and an evaluation set (50 normal/50 abnormal).

To create an appropriate experimental paradigm, only one EEG channel was selected for consideration. Examination of manual interpretation techniques practiced by experts revealed that the most promising channel to explore was the differential measurement T5-O1, which is part of the popular TCP montage [6]. This channel represents the difference between two electrodes located in the left temporal and occipital lobes. The spatial representation of this channel for a TCP montage is highlighted in Figure 2.

The first 60 seconds of each recording were used to extract signal features. The features were extracted through a standard cepstral coefficient-based approach that resembles the Mel Frequency Cepstral Coefficients (MFCCs) utilized in speech recognition [7]. Eight cepstral coefficients are used. These features were augmented with a differential energy term that accentuates the difference between quasi-periodic signals such as periodic lateralized epileptiform discharges (PLED) and background noise, bringing the dimension of the absolute feature vector to 9. First and second derivatives are added to the feature vector, bringing the total dimension to 27.

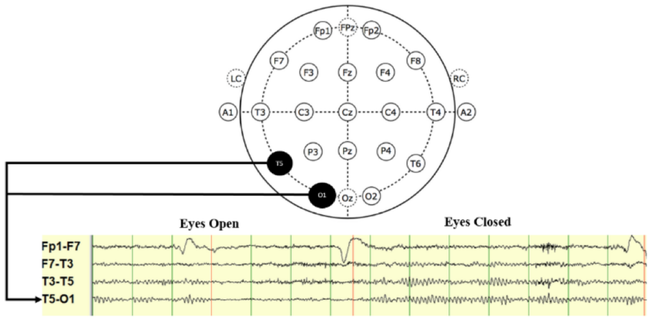

Figure 2. Emergence of the posterior dominant rythym (PDR) when the subject's eyes are closed. The spatial location of the channels used for classification, T5 and O1, are highlighted in the diagram.

A frame duration of 0.1 secs was used in the feature extraction process. The feature vectors from the first 60 secs of data were concatenated into a supervector of dimension 600x27=16,200. The dimensionality of the supervector was reduced using class-dependent Principal Components Analysis (PCA) in which we retained the N most significant eigenvectors of the covariance matrix [8] for each class.

Two standard algorithms were explored: k-Nearest Neighbor (kNN) [9] and Random Forest Ensemble Learning (RF) [10]. For kNN, class assignments were made by considering a majority vote of the k nearest neighbors. A class-specific Mahalanobis distance [9] was used in the analysis.

The specific RF algorithm used was based on a MATLAB implementation [11] of the algorithms described in [10]. An ensemble of trees $\{T_b\}_1^B$ was formed which produce an output classification given by:

$$\hat{C}_{rf}^{B}\,(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B \tag{1}$$

In essence, a class prediction $\hat{C}_b(x)$ for the bth tree is produced, and the final classification decision $\hat{C}_{rf}^{B}\,(x)$ is made in accordance to the majority of the classification results yielded by the ensemble of trees.

III. Experimental Results

The first parameter that needed to be tuned was the number of dimensions used for the PCA analysis. The original feature vector dimension of 1620 is obviously too large for our small dataset. There are several more sophisticated strategies that can be used to reduce its dimensionality including segmental averaging and a kernel-based rotation [12]. In this study we used a straightforward reduction in which we rank order the eigenvalues and discard the least significant eigenvectors [8].

Figure 3 demonstrates the performance of the RF algorithm as a function of the number of trees, N_t. It can be seen that performance does not improve significantly for $N_t > 20$. We selected $N_t = 50$ as a compromise between performance, complexity and computation time.

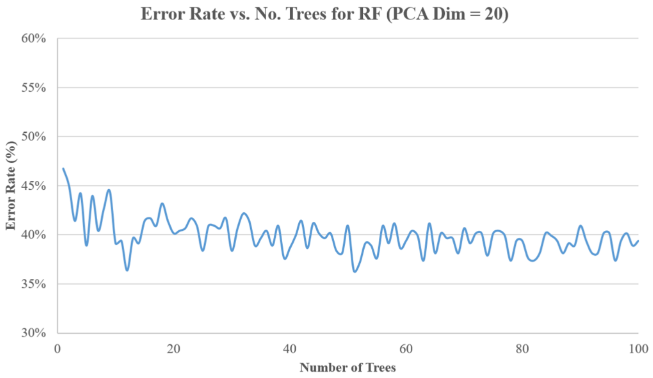

Figure 3. RF performance as a function of number of trees, N_t, is shown. Performance saturates for $N_t > 20$.

In Figure 4 we explore optimization of the number of output dimensions used in PCA for two algorithms: kNN with k = 1 and RF with N_t = 50. These plots are generated using a forced-choice paradigm in which one of the two classes is always chosen (rejecting both hypotheses is not an option). Both RF and kNN demonstrate that a PCA dimension of approximately 20 is adequate to obtain good performance. The first eigenvalue explains 99% of the variance, which is an indication that the features lack discriminating power.

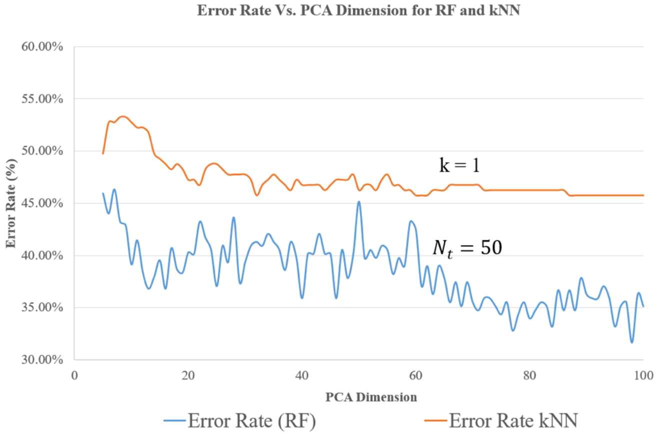

Figure 4. The forced-choice error rate for normal/abnormal classification is shown as a function of the number of PCA dimensions retained for RF and kNN.

Next, we evaluated performance as a function of the number of nearest neighbors in the kNN algorithm the PCA dimensions of 20 and 86. The results are shown in Figure 5. The performance of the system is best when k is in the range of 20 to 60. The data set is relatively small so we observe some amount of saturation in performance. We selected k = 20 for our operating point. Performance does not improve significantly beyond this value, and minimizing k reduces the computational requirements.

With these basic parameters now optimized, we explored which channel should be used for the analysis. The error rate as a function of the PCA dimension was studied for a value of k = 20 for a posterior temporal to occipital EEG channel (T5-O1) and a right frontal to central channel (F4-C4). Figure 6 presents these results. The T5-O1 channel is consistently better than F4-C4, which supports the clinical use of this differential channel.

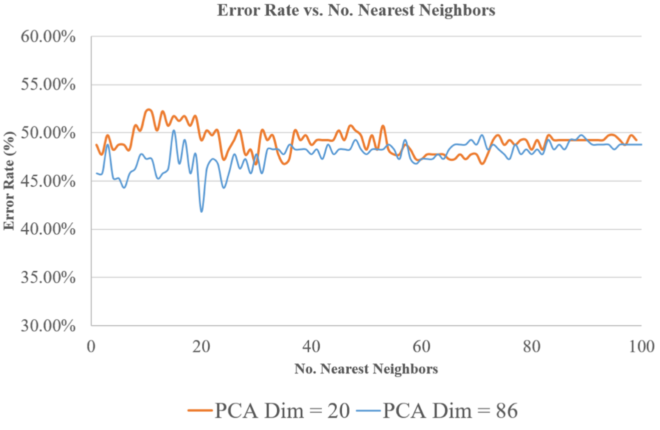

Figure 6. kNN performance as a function of k.

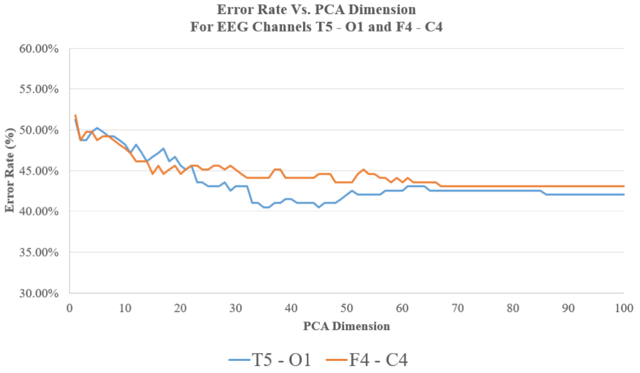

Figure 5. Performance of the system for a temporal to occipital (T5-O1) and a frontal to central (F4-C4) EEG channel. The performance for the T5-O1 channel was verified to be consistently better for PCA dimensions higher than 20.

Based on these results, we conducted additional searches for an optimal set of parameters for each system. In Table 1, we compare performance of two optimized systems to a baseline. The first system is random guessing based on priors. The second system is kNN with k = 20 and a PCA dimension of 86. The third system is RF with N_t = 50 and a PCA dimension of 86. In Table 2, we show a confusion matrix for the kNN system (the confusion matrix for RF is similar).

It is important to highlight the fact that for the tuning of each parameter, the operating point with the best performance was selected. In the cases where the performance of two or more different operating points was comparable, the point with better performance and less computational time was selected. For this reason, the number of trees for the RF algorithm, N_t, was selected to be 50 trees, while the kNN algorithm used k = 20.

Table 1 demonstrates that the tuned kNN and RF systems outperform random guessing based on priors, which is a promising outcome for these experiments. The balance of the normal/abnormal errors presented in Table 2, however, highlights the fact that there is a high confusion rate for normal EEGs. The dominant error is a normal EEG classified as abnormal. This could be explained by the presence of benign variants, or electroencephalographic patterns that resemble abnormalities, but do not qualify as events that would be of

No.	System Description	Error
1	Random Guessing	49.8%
2	kNN (k = 20)	41.8%
3	RF (N_t = 50)	31.7%

Table 1. A comparison of performance for our final three optimized systems is shown. kNN and RF perform significantly better than random guessing based on prior probabilities.

	Normal	Abnormal
Normal	50.5%	49.5%
Abnormal	34.0%	66.0%

Table 2. A confusion matrix generated for the best kNN system.

significance for the abnormal classification of a record. Also, we have not attempted to employ more sophisticated models of normal EEGs that include explicit models for events like artifacts and eye movements [7].

The computational time for training and evaluation for each algorithm is shown as a function of k and Nt, respectively, in Figure 7. kNN was considerably faster than the RF for training, but considerably slower for recognition, behavior that is explained by the nature of each algorithm. The behaviors presented in Figure 7 support the decisions for the tuning of each system parameter in both cases.

In a previous instance of this investigation, it was stated that parameters such as k and N_t were selected according to their performance and computational efficiency. For the Random Forest algorithm, for instance, the performance after 45 trees became comparable with the performance observed with greater number of trees, reaching optimal operation from N_t=48 to N_t=52. It is important to clarify that this performance was also achieved with systems with a greater number of trees, but that the system with 50 trees was selected because the point proved to be beneficial considering the tradeoff between performance and computational efficiency. The training of this final system took 183.02 seconds in total for the completion of the training and evaluation.

Contrary to how it was done with the RF algorithm, the optimal parameters for the kNN algorithm were selected through the consideration of the performance and the evaluation time. The training time through the kNN algorithm was relatively unaffected by the value of k, staying closely under 1 second for most values of k. The optimal parameters were then selected by taking into consideration the computational efficiency for the evaluation. The performance of the system for different values of k seemed to be comparable, showing the optimal performance when the PCA dimension was set to 86. The k value of 20 was then selected because the computational time for values of k below 30 was optimal. In this sense, the final system was tuned to have a value of k = 20. The total time required for the training and evaluation of this system was 2.14 seconds.

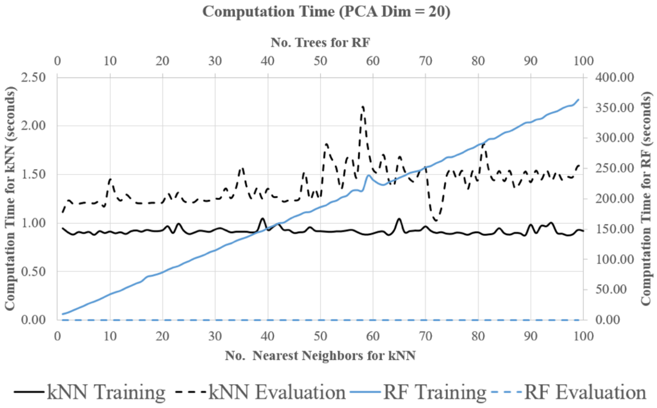

Figure 7. An analysis of the computational requirements for kNN and RF are shown. The computation time for the training through the RF algorithm is directly proportional to the number of trees, while the evaluation is very fast. The training through the kNN algorithm is close to 1 second for most values of k, and the evaluation becomes computationally heavier for values of k greater than 45, reaching a maximum of 2.2 when k = 58.

IV. Summary and Future Work

The present study has focused on the establishment of a proper experimental paradigm for the automated classification of normal/abnormal EEGs. A baseline experiment was presented that we hope will serve as a reference point for future studies. Two approaches, kNN and RF, were evaluated on features generated by using a PCA dimensionality reduction on the first 60 secs of EEG data. We have shown that the RF approach is better than the guessing based on priors, and resulted in an overall classification error rate of 31.7%. The system demonstrated better performance for the classification of abnormal records as abnormal, and had a higher confusion rate for normal files being identified as abnormal. Part of this behavior could be attributed to the benign variants that are often present in EEGs, such as Post Occipital Sharp Transients of Sleep (POSTs), which could potentially contribute to an erroneous classification.

EEG interpretation knowledge presented in [8], [9], [10] and [11] has been used in order to establish a system that resembles the common methods and techniques implemented by clinicians. Knowledge about the importance of the posterior dominant rhythm was used select the T5-O1 channel for processing. We verified that this channel appears to be rich in information for this task.

There are a number of obvious extensions of this work. First and foremost, we need to incorporate more temporal information into the process. This can be easily done building on the concepts presented in [7]. We can also incorporate more channels into the processing steps. Further, we can introduce more in sophisticated models for the normal class label, which essentially functions as a universal background model [13]. Finally, we can detect additional features, such as those described Figure 1, and incorporate this information into the multi-level processing scheme described in [7].

Note that the data used on this study is publicly available at *www.nedcdata.org*. It is a subset of the TUH EEG Corpus which is also available at the same URL.

V. Acknowledgements

The primary funder of this research was the QED Proof of Concept program of the University City Science Center, Research reported in this publication was also supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U01HG008468 and the National Science Foundation through Major Research Instrumentation Grant No. CNS-09-58854. The TUH EEG database work was funded by the Defense Advanced Research Projects Agency (DARPA) MTO under the auspices of Dr. Doug Weber through the Contract No. D13AP00065, Temple University's College of Engineering and Temple University's Office of the Senior Vice-Provost for Research. Finally, we are also grateful to Dr. Mercedes Jacobson, Dr. Steven Tobochnik and David Jungries of the Temple University School of Medicine for their assistance in developing the classification paradigm used in this study and for preparing the manually annotated data.

References

[1] F. Fahoum, R. Lopes, F. Pittau, F. Dubeau, and J. Gotman, “Widespread epileptic networks in focal epilepsies: EEG-fMRI study,” *Epilepsia*, vol. 53, no. 9, pp. 1618–1627, Sep. 2012.

[2] H. Azuma, S. Hori, M. Nakanishi, S. Fujimoto, N. Ichikawa, and T. A. Furukawa, “An intervention to improve the interrater reliability of clinical EEG interpretations,” *Psychiatry Clin. Neurosci.*, vol. 57, no. 5, pp. 485–489, Oct. 2003.

[3] S. Smith, “EEG in the diagnosis, classification, and management of patients with epilepsy,” *J. Neurol. Neurosurg. Psychiatry*, vol. 76, no. Suppl 2, pp. ii2–ii7, Jun. 2005.

[4] J. S. Ebersole and T. A. Pedley, Current practice of clinical electroencephalography, 4th ed. Philadelphia, Pennsylvania, USA: Wolters Kluwer, 2014.

[5] A. Harati, S. Lopez, I. Obeid, M. Jacobson, S. Tobochnik, and J. Picone, “THE TUH EEG CORPUS: A Big Data Resource for Automated EEG Interpretation,” in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 2014, pp. 1–5.

[6] A. C. N. Society, “Guideline 6: A Proposal for Standard Montages to Be Used in Clinical EEG [White Paper]. Retrieved from *http://www.acns.org/pdf/guidelines/Guideline-6.pdf*, 2006.

[7] A. Harati, M. Golmohammadi, S. Lopez, I. Obeid, and J. Picone, “Improved EEG Event Classification Using Differential Energy,” in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 2015, pp. 1–4.

[8] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York City, New York, USA: Springer-Verlag, 2002.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. New York City, New York, USA: John Wiley & Sons, 2003.

[10] L. Breiman, J. Friedman, R. A. Olshen, and C. Stone, *Classification and Regression Trees*, 1st ed. Boca Raton, Florida, USA: Chapman and Hall/CRC, 1984.

[11] Mathworks, “TreeBagger,” *Statistics and Machine Learning Toolbox*, 2015. [Online]. Available: *http://www.mathworks.com/help/stats/treebagger.html*. [Accessed: 18-Oct-2015].

[12] A. Ganapathiraju, J. Hamaker, and J. Picone, “Applications of Support Vector Machines to Speech Recognition,” *IEEE Trans. Speech Audio Process.*, 2002.

[13] Reynolds, D., & Campbell, W. (2008). Text-Independent Speaker Recognition. In *Springer Handbook of Speech Processing* (1st ed., p. 1176). Berlin, Germany: Springer.