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}@temple.edu, joseph.picone@gmail.com

***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., evolution). 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 on EEG interpretation. We have been developing an interpretation system, AutoEEGTM, 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 41.79% detection error rate while RF an error rate of 31.66%. 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 the normal classification.**

# 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 in 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 for the EEG interpretation 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 subjects to more than one EEG prior to a diagnosis.

In this sense, the automated classification of EEGs as normal or abnormal records 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 the specialized neurologists save time in their daily EEG interpretation routine, easing some of the service pressures that arise from the increasing demand of EEGs [3].

The classification of an EEG record as normal or abnormal is an assessment that is made through the observation and examination of certain characteristics, or lack thereof that contribute to the normality of the record. 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 6-7 Hz activity present in the frontal or frontocentral regions of the brain.

Neurologists can usually make this determination by examining the first few minutes of a recording. Hence, in this baseline study, we will focus on this problem in an attempt to calibrate the difficulty of the machine learning problem.

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, is taken in this study as the main decisive feature for the establishment of a normal/abnormal classification baseline, mainly because of its distinctive and prevalent characteristic in a normal EEG. Additionally, the fact that this feature appears mostly occipitally provides a logical advantage for the purpose of the formation of an experimental paradigm, because it allows to select one occipital channel from each recording to make a classification.

# 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 set an appropriate experimental paradigm in place, only one EEG channel was selected to be taken into 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.

A frame duration of 0.1 secs was used in the feature extraction process. The first 60 secs of data was concatenated into a supervector of dimension 60x27=1620. The time and space complexity inherent to the dimensionality of the computed feature vectors was reduced through the representation of the data in a lower dimensional space. This was achieved through the computation of the residuals obtained from the retention of the principal components of the concatenated matrix comprised by the feature vectors [8].

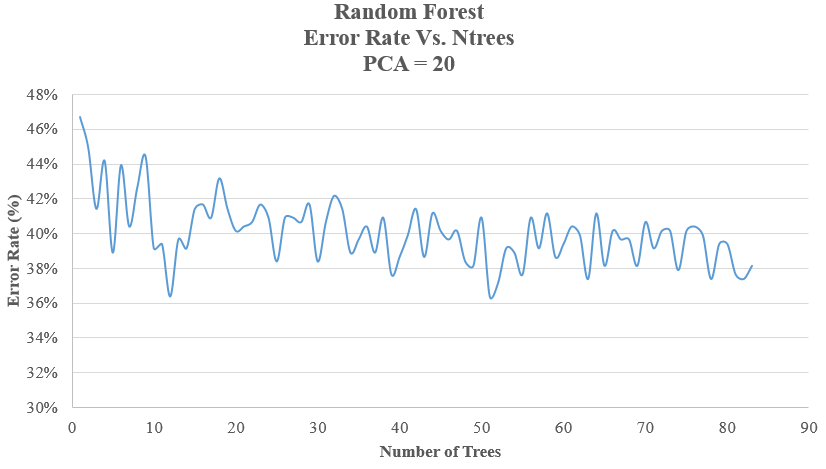


Figure 1. Random Forest as a function of number of trees. The performance of all the systems anallyzed after 20 trees seem to be comparable.

Figure 1. 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.

Two standard algorithms were explored: k-nearest neighbor (kNN) [9] and random forests (RF) [10]. The kNN approach used the reduced dimension PCA output for its input. Models for each class were built by averaging feature vectors for each class. These vectors were normalized using a class-specific covariance matrix. Class assignments were made by considering a majority vote of the k nearest neighbors. A 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 was formed which produce an output classification given by:

(1)

In essence, a class prediction for the bth tree is produced, and the final classification decision is made in accordance to the majority of the classification results yielded by the ensemble of trees.

# 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, popular with PCA, in which we rank order the eigenvalues and discard the least significant eigenvectors [8].

Figure 1 shows the performance of the RF algorithm as a function of the number of trees. It can be seen that the performance of the systems higher than 20 trees are comparable to each other. In this sense, taking the performance and the computational time for the classification into account, a number of 50 trees was chosen for the rest of the experiments.

In Figure 2 we explore performance as a function of the PCA dimensions for two algorithms: kNN with k = 1 and RF with NTrees = 50. We also show the percent of the variance explained by the PCA dimension. 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).

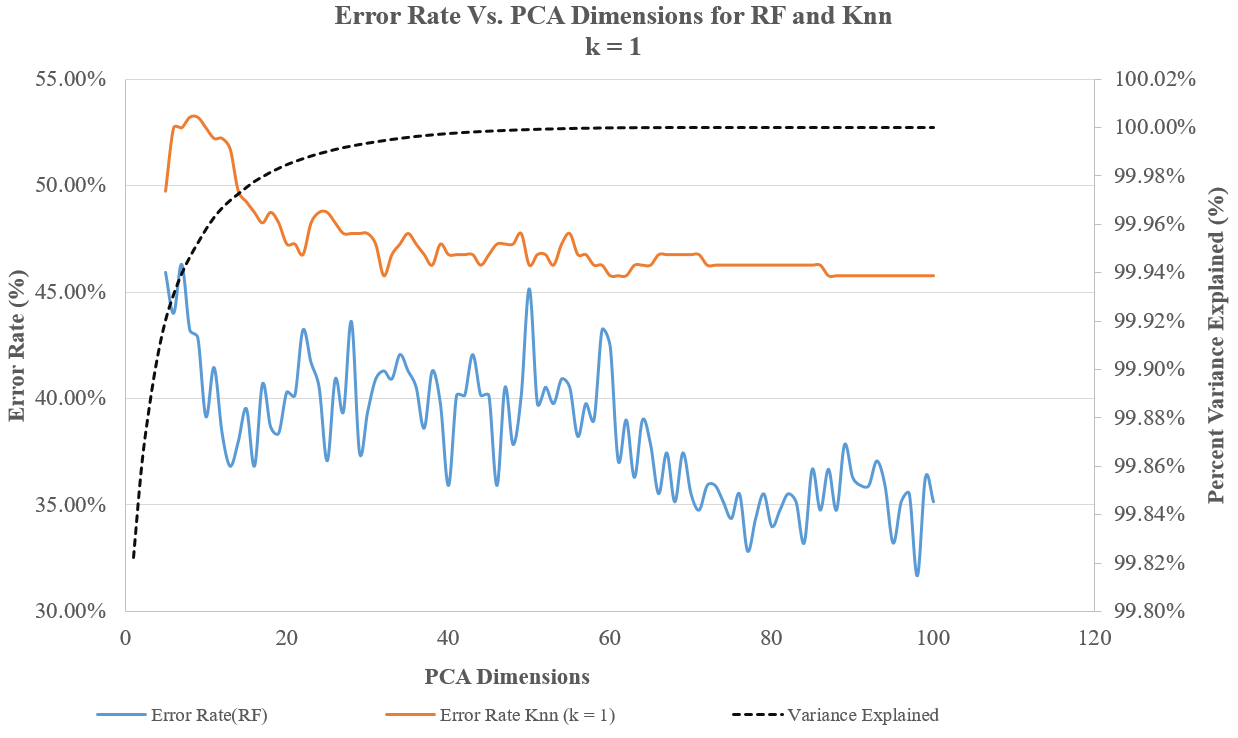
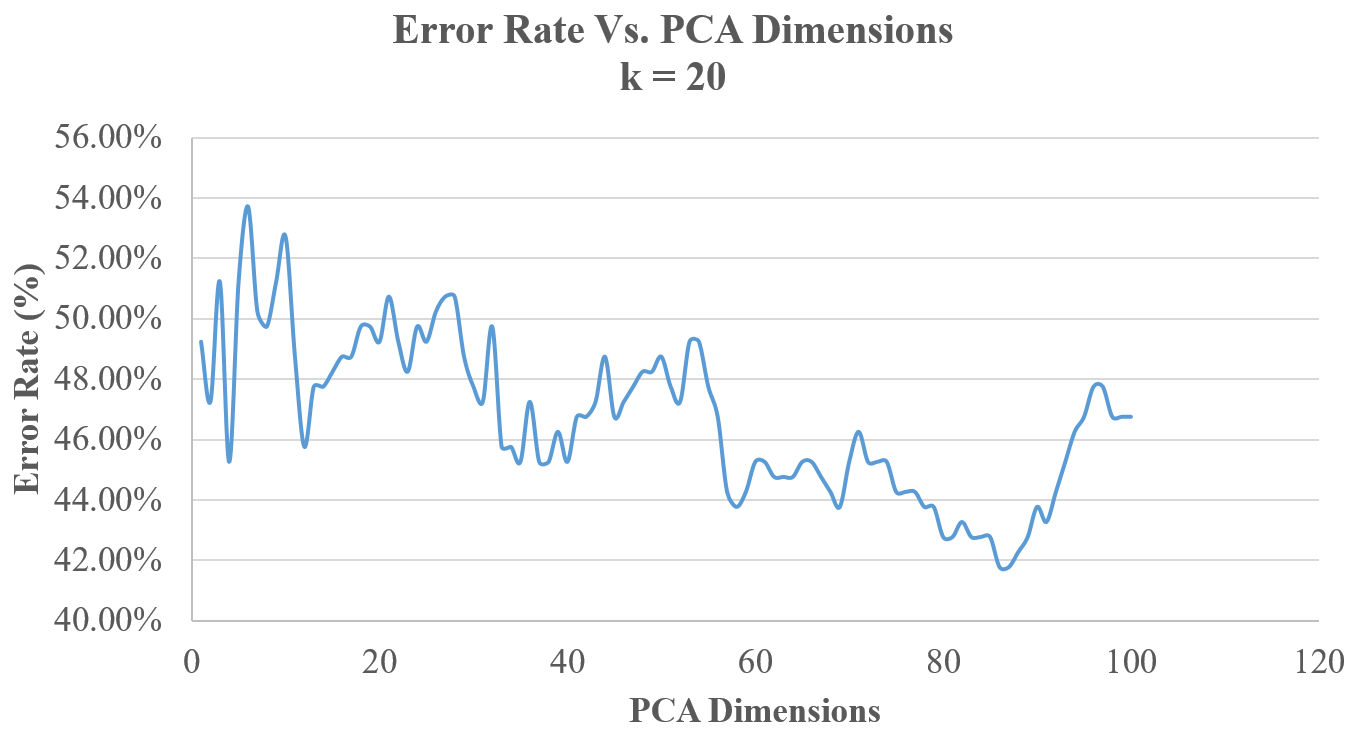
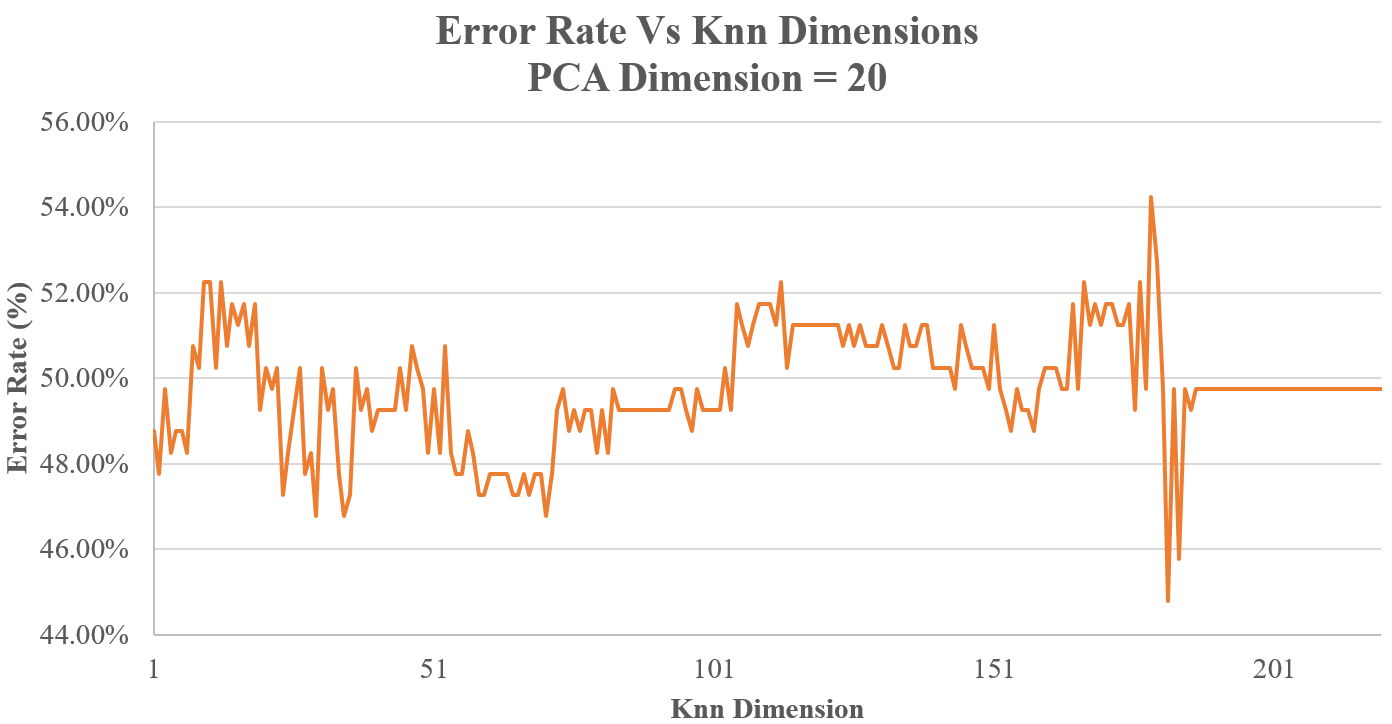


Figure 2. The forced-choice error rate for normal/abnormal classification is shown as a function of the number of PCA dimensions retained. We compare kNN (k=1), RF and the percent variance explained. In this plot, it is possible to see that the RF algorythm exhibits considerably better performance than kNN and random guessing based on priors for all PCA dimensions. For PCA dimensions higher than 20, the performance of both systems was better than random guessing based on priors. It can be seen that 99.82% of the variance is explained by the first principal component.

Figure 4. System performance as a function of PCA for k = 20. It can be observed that tthis system achieves an error rate of 41.79% when the PCA dimension is 86.

Next, we evaluated performance as a function of the number of nearest neighbors in the kNN algorithm for a fixed PCA dimension of 20. The results are shown in Figure 3. The performance of the system is the best when k is in the range of 20 to 60 with some statistical variability. The data set is relatively small so we observe some amount of saturation in performance, especially after k is greater than 200. Based on this analysis, we set k=20 to be our optimal operating point.

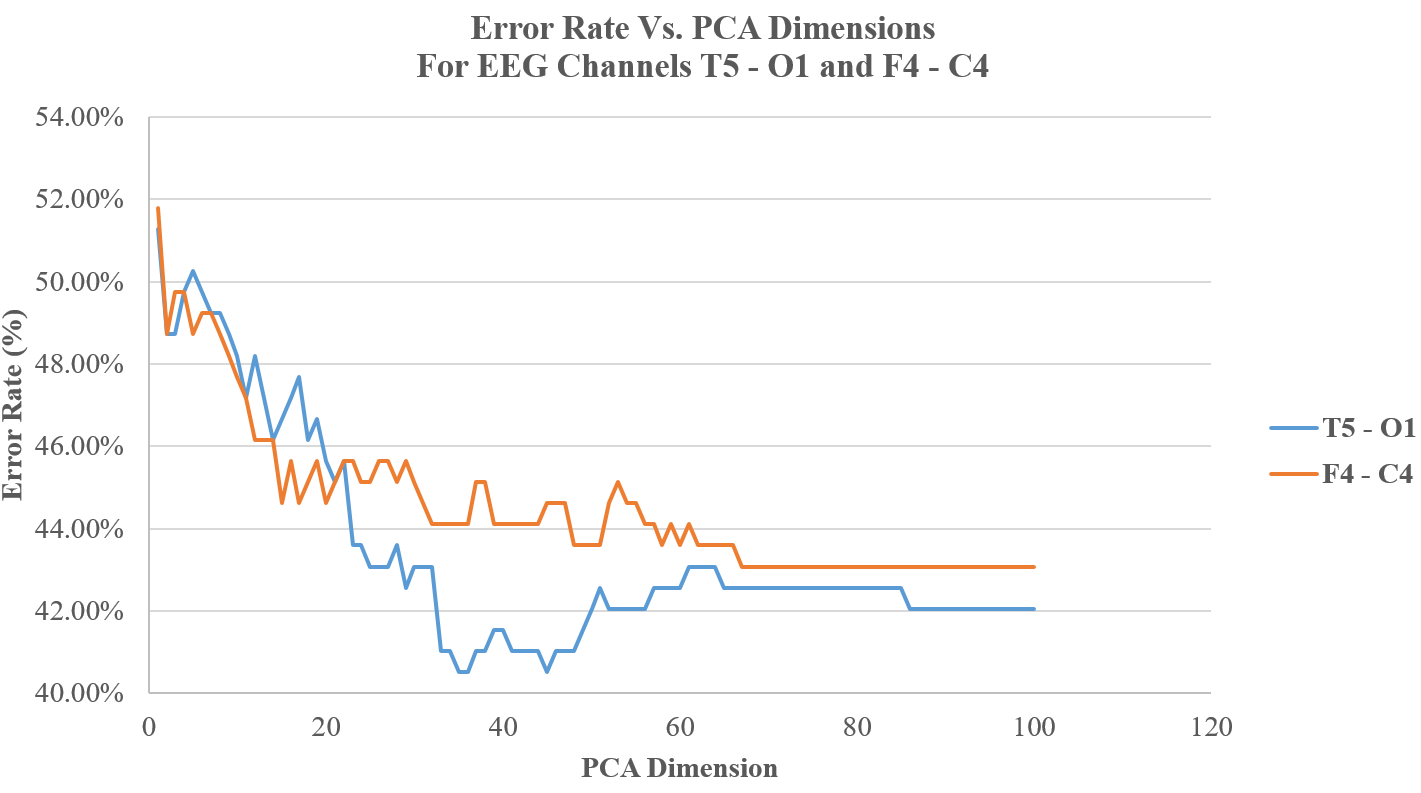


1. Figure 3. Performance of kNN is shown as a function of k. It is possible to see that the performances is the best when k is between 20 and 60. An increase in the error rate, followed by a saturation in performance can be observed as k kincreases past the mentioned interval.

The lowest k for the best operating interval was chosen in order to reduce the computational impact related to the number of the nearest neighbors that had to be analyzed in order to determine most suitable classification decision. The performance of the system when k is set to 20 as a function of PCA dimensions can be seen in Figure 4.

The effect of the dimensionality in the PCA can be seen in Figure 4, where the classification error shown by the curve tends to decrease, with some statistical variation, as the dimensions get higher.

Another important aspect of the problem that was experimentally analyzed, was the performance of the system for two different EEG channels. The error rate as a function of PCA dimensions for a value of k = 1 was studied for a posterior temporal to occipital EEG channel (T5-O1) and a right frontal to central channel (F4-C4). Figure 5 shows the results obtained in this experiment.



1. 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 better for all operation points with PCA dimensions higher than 20. As the dimensionality increases, both performances saturate, with the temporal to occipital channel being the best performance.

The significance of the channel analyzed over the performance of the system is highlighted in Figure 5, where it is possible to see that the temporal to occipital channel’s analysis shows a better performance than the frontal to central channel. It can be noted that the observations made in this experiment correlate with the information that we have learned from physicians about their reliance on the occipital channels for the normal and abnormal classification of EEGs.

Based on these optimizations, in Table 1 we show the confusion matrix calculated from our best system for the kNN algorithm, with a k = 20 and a PCA = 86. Table 2, the performance of our three final systems: (1) random guessing based on prior probabilities, (2) kNN with k=20, and (3) RF with Tree = 25.

|  |  |  |
| --- | --- | --- |
|  | **Normal** | **Abnormal** |
| **Normal** | **50.49%** | **49.50%** |
| **Abnormal** | **34.00%** | **66.00%** |

Table 1. Confusion matrix generated with the results of our best kNN system. It is evident from the table that there is a high confusion rate in the classification of normal files.

|  |  |  |
| --- | --- | --- |
| **No.** | **System Description** | **Error** |
| **1** | **Random Guessing** | **49.75%** |
| **2** | **kNN (k = 20)** | **41.79%** |
| **3** | **RF (Ntrees = 50)** | **31.66%** |

Table 2. A comparison of performance for our final three systems is shown. kNN and RF perform significantly better than random guessing based on prior probabilities.

Table 2 shows that the tuned kNN and RF systems outperform the random guessing based on priors, which is a promising outcome of the experiments. The balance of the normal/abnormal errors presented in Table 1, however, highlights the fact that there is a high confusion rate of normal EEGs as abnormal EEGs, fact that 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 significance for the abnormal classification of a record.

# Summary and Future Work

The present study has focused in the establishment of a proper experimental paradigm for the automated classification of normal/abnormal EEGs. A baseline experiment has been set for reference of future studies in the issue. The classification decisions were made through the random forest ensemble learning method and the results were later compared to the performance obtained from the application of a kNN algorithm and the guessing based on prior information. The experiments conducted have shown that the random forest approach is better than the guessing based on priors, which is also outperformed by the tuned kNN system.

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 the specialized neurologists but increases the interrater agreement through the introduction of this automated classification method. Essentially, the knowledge about the posterior dominant rhythm and its particular characteristics was used for the selection of a significant EEG channel for this particular study.

The experimental results made evident that the channels taken into account for the classification are of great significance for the performance of the systems. As it had been hypothesized from the discussion with specialized neurologists, the system showed better performance for the classification of abnormal records as abnormal, and had a higher confusion rate of normal files with abnormal ones. 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.

As it was established, state of the art machine learning techniques will be applied for the improvement of the system through the introduction of new labels that will augment the amount of information that is used in order to make a classification decision and more sophisticated temporal modeling techniques, such as Hidden Markov Models. This study shows that time should be invested in the analysis and extraction of features that are more meaningful and adequate to this particular classification problem. Additionally, it is important to take benign variables into account when the construction of a more sophisticated classification system is in progress, which could be done through the labeling of events characterized by physicians as events that would not contribute to an abnormal classification.

# 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.