AUTOMATED INTERPRETATION OF ABNORMAL
ADULT ELECTROENCEPHALOGRAMS

A Thesis
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
MASTER OF SCIENCE
ELECTRICAL ENGINEERING

by
Silvia López de Diego
August, 2017

Thesis Approvals:

Dr. Joseph Picone, Advisory Chair, Department of ECE
Dr. Iyad Obeid, TU Department of ECE
Dr. Pallavi Chitturi, TU Department of Statistics

ABSTRACT

Interpretation of electroencephalograms (EEGs) is a process that is still dependent on the subjective analysis of the examiner. The interrater agreement, even for relevant clinical events such as seizures, can be low. For instance, the differences between interictal, ictal, and post-ictal EEGs can be quite subtle. Before making such low-level interpretations of the signals, neurologists often classify EEG signals as either normal or abnormal. Even though the characteristics of a normal EEG are well defined, there are some factors, such as benign variants, that complicate this decision. However, neurologists can make this classification accurately by only examining the initial portion of the signal. Therefore, in this thesis, we explore the hypothesis that high performance machine classification of an EEG signal as abnormal can approach human performance using only the first few minutes of an EEG recording.

The goal of this thesis is to establish a baseline for automated classification of abnormal adult EEGs using state of the art machine learning algorithms and a big data resource – The TUH EEG Corpus. A demographically balanced subset of the corpus was used to evaluate performance of the systems. The data was partitioned into a training set (1,387 normal and 1,398 abnormal files), and an evaluation set (150 normal and 130 abnormal files). A system based on hidden Markov Models (HMMs) achieved an error rate of 26.1%. The addition of a Stacked Denoising Autoencoder (SdA) post-processing step (HMM-SdA) further decreased the error rate to 24.6%. The overall best result (21.2% error rate) was achieved by a deep learning system that combined a Convolutional Neural Network and a Multilayer Perceptron (CNN-MLP). Even though the performance of our algorithm still lags human performance, which approaches a 1% error rate for this task, we have established an experimental paradigm that can be used to explore this application and have demonstrated a promising baseline using state of the art deep learning technology.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The recording of electrical activity along the scalp, known as electroencephalography (EEG), has been widely used for the diagnosis and management of conditions such as sleep disorders and epilepsy during the past 30 years. Despite the emergence of new technologies, such as Magnetic Resonance Imaging (MRI), the noninvasive nature and relative low cost of an EEG make this technique a popular choice as a diagnostics tool among physicians (Smith, 2005). A typical routine outpatient EEG has a duration of about 20 minutes. This duration is not always adequate to record ictal (or interictal activity) in patients with seizure disorders. As a matter of fact, only 50% of patients with epilepsy show interictal epileptiform discharges (IED) in their first recording (Smith, 2005). The diagnosis and characterization of epilepsy, which is a life-altering diagnosis, usually requires multiple EEG sessions and/or a long-term monitoring (LTM) recording. LTMs are typically many hours to several days in duration and are administered as an in-patient service, which makes them extremely expensive.

EEG records are manually interpreted by board certified physicians. Because this is a time-consuming process, the lag time in reading an EEG can range from hours to weeks. Additionally, the interpretation of an EEGs depends heavily on the subjective judgement of the reader, which can lead to missed events or misdiagnosis of the patient (Azuma et al., 2003). Introducing a certain level of automation to the EEG interpretation task could potentially serve as an aid for the neurologists to accelerate the reading process and ease the pressure that results from a heavy case load.

The EEG recording workflow for a scalp EEG, which is the most common form of an EEG administered today, is summarized in Figure 1. A typical recording involves the placement of the
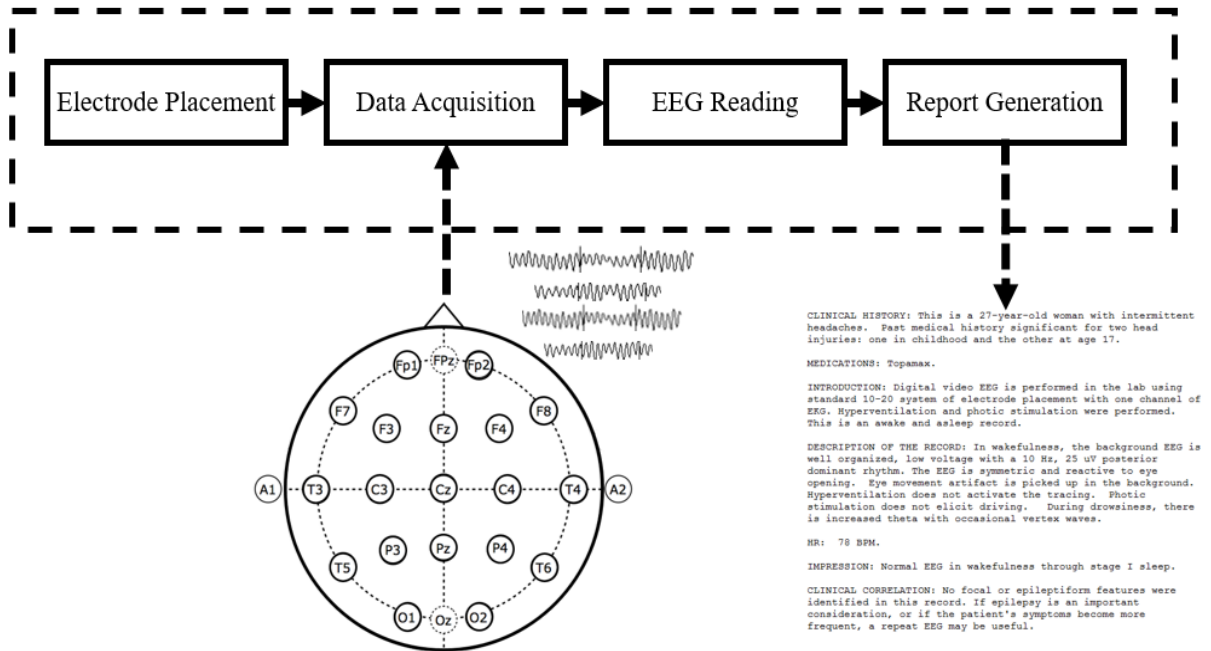
Figure 1. Summary of the common steps that are followed for a clinical EEG examination

electrodes on the patient's scalp by a technician, the acquisition of the EEG data, the interpretation

of the signals by a certified neurologist and the generation of the report that is presented to the

patient (Harati et al., 2014). An EEG report contains a combination of the history, medications,

description of the record and interesting findings. One portion of the report, however, contains the

impression of the record, which shows whether the EEG is normal or abnormal given the EEG

activity recorded in the session. Automatic determination of an abnormal EEG using machine

learning techniques is the focus of this work.

The medical report that is produced for each EEG session describes the record, the

recording conditions and summarizes the findings. One decision that is also shown in the report is

whether the characteristics of the EEG was found to be within the normal limits for patients in a

similar group of age and gender. The main objective for this study is to utilize machine learning

techniques to generate this decision automatically. If the proposed technology reaches clinically

acceptable performance, it could potentially serve as an aid to neurologists and reduce the lag time between EEG recording and reporting, thereby establishing a more efficient workflow.

1.1   The Normal Adult EEG

The EEG interpretation task consists of two parts: the analysis of the EEG background signal and recognition of transients (Finnigan & van Putten, 2013). The patterns present in the background signal represent the general characteristics of an EEG, which include the features that neurologists observe when making a normal/abnormal decision about the record. Some prominent examples of these background patterns are the posterior dominant rhythm and the frequency distributions of the signals throughout the scalp (Lodder & van Putten, 2013). The transient patterns, on the other hand, refer to rarer events that include pathological and physiological waveforms, such as spikes and sharp waves discharges.

Characterization of a normal adult EEG has been based on a specific description of the background patterns and the presence — or lack thereof — of certain transient waveforms given the patient's state of consciousness (awake vs. drowsy vs. comatose). The main background characteristics of a normal adult EEG can be summarized as follows (Ebersole & Pedley, 2014):

1. *Reactivity:* Refers to the response to certain physiological changes. These changes could be due to eye opening and closing, sensory stimulation, etc.

2. *Alpha Rhythm:* This rhythm is the starting point for the visual analysis of EEGs. The characteristics of this feature play an important role in the normal/abnormal classification of the EEG. Alpha waves originate (predominantly) in the occipital lobe and are between 8-13 Hz in frequency and 15 to 45 μV in amplitude.

3. *Mu Rhythm:* It is a central rhythm of frequencies between 8 to 10 Hz with amplitudes comparable to the alpha rhythm. This rhythm is suppressed unilaterally by the movement

3

of the opposite limb (e.g.: left hemisphere mu rhythm is suppressed when the subject performs a motor action with a right-sided limb). This rhythm, however, is also suppressed by conditions such as fatigue, somatosensory and sensorimotor stimulation. In this sense, the Mu rhythm is not always detectable.

4. *Beta Activity:* A rhythm that contains frequencies between 18-25 Hz, 14-16 Hz and 35-40 Hz, with amplitudes between 5 and 20 mV. It is important to note, however, that it is rare to see activity higher than 25 Hz in frequency spectrum on scalp EEGs.

5. *Theta Activity:* Normal adults tend to show traces of less than 15 μV 6-7 Hz activity in the frontal and frontocentral regions and occasionally in the midline central region. This rhythm, called theta activity, usually becomes sustained and higher in voltage with the onset of drowsiness.

These features provide a description of the characteristics that are systematically observed by neurologists when evaluating an EEG. Factors such as the state and the age of the patient are also important considerations that may alter the signals described above. These characteristics are mainly observed in normal adult EEGs (Ebersole & Pedley, 2014).

1.2   Visual Analysis of EEGs

The previous section described the general features that characterize a normal EEG. The presence of these features does not necessarily guarantee the normality of the record. Neurologists analyze records by evaluating the background signal and determining whether the patient presents normal characteristics according to his or her state. If the patient did not present abnormal transients during the recording, and the background EEG was within normal limits, the recording is considered normal. A decision tree for the evaluation of the normality of an EEG record is presented in Figure 2 (Lopez et al., 2015). The analysis of the background is decomposed into a

4
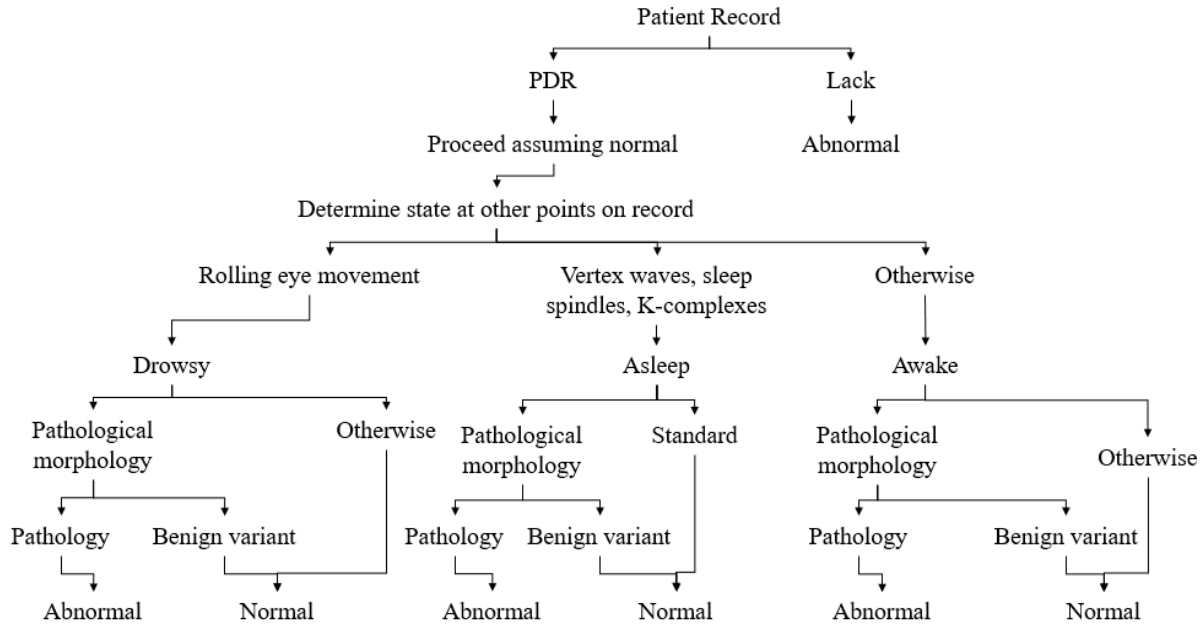
Figure 2. A decision tree demonstrating the logic used to classify an abnormal EEG

series of steps that account for all relevant characteristics in a systematic way. The analysis steps that ultimately lead to a decision about the normality of the record involve the observation of the following characteristics: frequency, voltage, waveform, regulation (e.g., the alpha rhythm should not vary more than ±0.5 Hz), locus, reactivity and interhemispheric coherence (Ebersole & Pedley, 2014).

The systematic visual analysis of EEGs usually starts with the evaluation of the occipital alpha rhythm, also called the Posterior Dominant Rhythm (PDR). This activity emerges in the occipital region when the eyes are closed, and fades as the patients enter a state of drowsiness. In this sense, the evaluation of the reactivity for the emergence of the PDR is one of the main features used to assess the normality of the record (Ebersole & Pedley, 2014). Throughout this study, domain knowledge like this from clinical neurology is integrated into a machine learning system to achieve a high-performance classification system.

1.3   Thesis Overview

A generalized algorithm or method for the classification of clinical abnormal EEGs is a task that has not yet been explored. Determination of normality is the first step in automatic interpretation of an EEG (Picone & Obeid, 2017) and can be used to reduce the false alarm rate on tasks such as seizure detection. High false alarm rates are a critical reason automated technology is not used in clinical settings (Scheuer et al., 2017). While some work (Peker, 2015; Fernandez-Varela, 2017) has been presented on the identification of EEG abnormalities specific to certain pathological or physiological conditions, the study of the general background EEG as a resource for the classification of normal and abnormal records has not been investigated. For instance, classification of athletes with residual functional deficits after the occurrence of a concussion has been attempted using EEG signals and Support Vector Machines (SVMs) (Cao et al., 2008). This study, however, did not rely on clinical EEG data, and the classifier was designed to recognize a very specific condition.

Hence, a major goal in this work was to establish baseline performance for the classification of abnormal clinical EEG records. A variety of machine learning approaches, introduced in Chapter 2, were developed and evaluated. Two important non-parametric algorithms, k-Nearest Neighbor (KNN) (Duda et al., 2001) and Random Forest Ensemble Learning (RF) (Breiman, 2001) were used to establish baseline performance. These extremely popular and robust approaches are a good way to assess the inherent difficulty of a task. Hidden Markov Models (HMM) (Picone, 1990), a similarly important parametric approach, were also utilized for classification and comparison with the baselines.

In addition to these baseline systems, we explored several popular deep learning approaches. Deep learning has, of course, attracted a lot of attention in recent years because state of the art

performance has been achieved on a broad range of problems (He et al., 2016; Saon et al., 2016; Bar et al., 2015). We introduce a hybrid HMM/deep learning system that postprocesses HMM outputs with a Stacked Denoising Autoencoder (SdA) (Vincent & Larochelle, 2010) and an end-to-end deep learning system based on a Convolutional Neural Network (CNN) (Goodfellow et al., 2017). In Chapter 2, overviews of our baseline KNN, RF and HMM systems are presented in the context of a sequential decoding problem. In addition, Chapter 2 includes a description of the application of CNN to sequential decoding problems such as speech recognition. The advantages that these networks offer, and the ways in which these advantages can be leveraged in an EEG decoding task are discussed.

In Chapter 3, we introduce the data used to develop and evaluate the technology presented in this study, along with a description of the experimental design. The development of a significant database for this task is an extremely important part of any machine learning research and one of the major contributions of this work. This data has been publicly released at *https://www.isip.piconepress.com/projects/tuh_eeg/* (Obeid & Picone, 2016). In Chapter 4, the results of our experiments are discussed, and performance is compared to human performance. Finally, Chapter 5 offers some concluding thoughts about future research to improve the performance of our systems.

CHAPTER 2

CLASSIFICATION OF SEQUENTIAL DATA

Electroencephalography signals, like speech signals, are the product of a physiological process that unfolds in time. Figure 3, for instance, shows the temporal evolution of a seizure in one EEG channel and its respective spectrogram. In this sense, machine learning approaches that treat the observations in the data as independent and identically distributed (i.i.d.) would not successfully exploit the sequential nature of the data (Bishop, 2011). Modeling the temporal evolution of the frequency spectrum of an EEG signal suggests HMMs would be a promising baseline approach for classifying abnormal EEGs. In this chapter, we introduce a variety of popular machine learning techniques that are capable of modeling this evolution. The theoretical explanations presented here follow that in Duda et al. (2001), Rabiner (1989) and Bishop (2011).

2.1   Nonparametric Models: k-Nearest Neighbor (kNN) and Random Forest (RF)

Before starting to discuss the specific modeling of sequential data with HMMs and deep learning, it is important to consider two classic non-parametric techniques that we used for an exploratory study of the subject: k-Nearest Neighbor (kNN) (Duda et al., 2003) and Random Forest (RF) (Breiman, 2001). These techniques are both well understood and have been widely adopted in different areas of pattern recognition (Keysers et al., 2007; Chu et al, 2015). Despite their antiquity, these models are still useful for pattern recognition problems in which the distributions that generate the data are difficult to model.
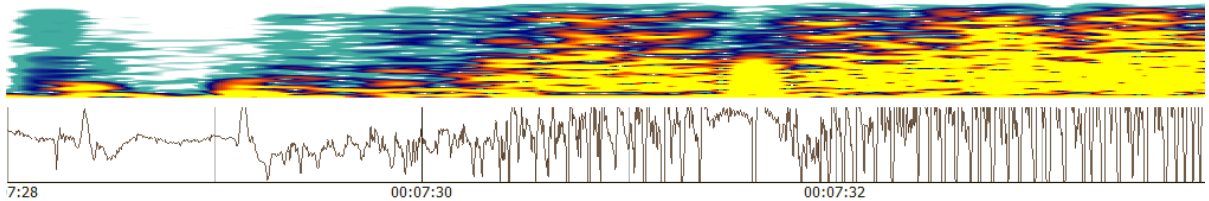
Figure 3. Temporal Evolution of a seizure in the T4-A2 channel of an EEG. The top of the figure shows the spectrogram of the signal, while the bottom panel shows the signal in the time domain.

One of the oldest and simplest pattern recognition algorithms is kNN. When combined with other techniques such as neural networks or prior knowledge, however, models that rely on kNN can achieve competitive results (Weinberger & Saul, 2009; Belongie et al., 2002). Explained in a simple way, kNN classifies a given data sample by considering the k-nearest neighboring samples and selecting the label that predominates in the considered window. Traditionally, in the absence of prior knowledge about the data, a simple Euclidean distance measure between the input vectors is used to calculate the distances between data samples, but the distance metrics are often adapted to the task being solved. Some systems, such as (Chopra et al., 2005), additionally learn a similarity metric from the data, which significantly improves the classification performance.

Another well-known machine learning technique that is still being successfully used for a variety of problems, such as gene selection and classification (Díaz-Uriarte, 2006), is RF. RFs are different from standard trees in that the nodes are split using a random set of samples specifically selected at that node, rather than the best split among all variables. In accordance with the RF algorithm presented by Breiman (2001), an ensemble of trees is formed in order to produce a class prediction. A final classification decision is then made through the consideration of a majority vote yielded by each ensemble of trees. RFs are an excellent choice for a baseline system because they are robust to overfitting and perform well across a wide range of applications. Some of the preliminary studies in our work rely on standard kNN and RF techniques to evaluate the abnormal EEG problem before considering HMMs and Deep learning approaches.

2.2   Hidden Markov Models

Consider a sequence of states where the state at a time t is denoted as $\omega(t)$. The description of the model for a specific sequence ω^T (where T represents the length of the sequence) is then given by Eq. 1:

$$P(\omega_j(t+1)|\omega_i(t)) = a_{ij},\tag{1}$$

where a_{ij} represents a transition probability, or the probability of being in state ω_j at $t+1$ given that the state at t is ω_i (Duda et al., 2001). The state at step $t+1$ in a first-order Markov model is a function that only depends on t. Higher order Markov chains allow to consider states at earlier steps. In the model in Eq. 1 which is referred to as an observable Markov model, each step corresponds to an observable event. Figure 4 shows an illustration of a three-state Markov model, with its respective states ω_i represented by nodes and the transition probabilities a_{ij} represented by links. Hidden Markov models can be considered an extension, or augmentation, of the models that have been described to this point. In fact, in the case of HMMs, the visible observations $v(t)$ are given by a probabilistic function of the state (Rabiner, 1989). In this augmented model, pictorially represented in Figure 5, the assumption that at every single time t the system is at state $\omega(t)$ is kept. However, for HMMs, the assumption that the system also emits a visible observation or symbol $v(t)$ is also made. In this way (assuming a discrete symbol is emitted at each state), a probability of emitting a specific visible state $v_k(t)$ is given by Eq. 2 (Duda et al., 2001):

$$P\left(v_k(t)|\omega_j(t)\right) = b_{jk}.\tag{2}$$

In speech recognition, each spoken word w is decomposed into a sequence of K_w sounds (or base phones), which have pronunciation sequences $q_{1:K_w}^{(w)} = q_1, \dots, q_{K_w}$. The likelihood of a word w given an acoustic feature vector Y is given by Eq. 3:
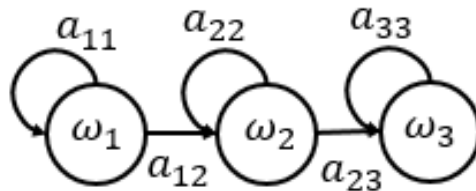
Figure 4. Example of a basic Markov model with states ω_i and transition probabilities a_{ij}

10
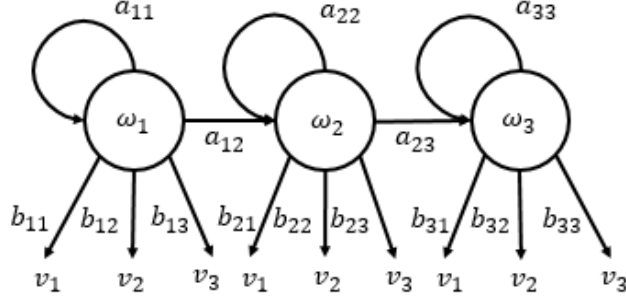
Figure 5. Example of a Hidden Markov model with states ω_i, transition probabilities a_{ij}, emission probabilities b_{jk} and visible stated v_k

$$\hat{w} = \underset{w}{\mathrm{argmax}}\{P(w|Y)\}\,. \tag{3}$$

Given the difficulty in modeling $P(w|Y)$, Bayes Rule can be used to transform this equation into the equation specified by Eq. 4:

$$\hat{w} = \underset{w}{\mathrm{argmax}}\{P(Y|w)P(w)\}\,, \tag{4}$$

where $P(Y|w)$ represents the acoustic model and $P(w)$ represents the language model (Gales & Young, 2007). This is much easier to model in practice.

If the decoding of the word "bat" is considered, for example, each of the valid pronunciations for the phones that comprise the word (/b/, /ae/ and /t/) would be represented by a continuous density HMM of the form shown by Figure 6 (Gales & Young, 2007), and the likelihood $P(Y|w)$ would be given by Eq. 5:
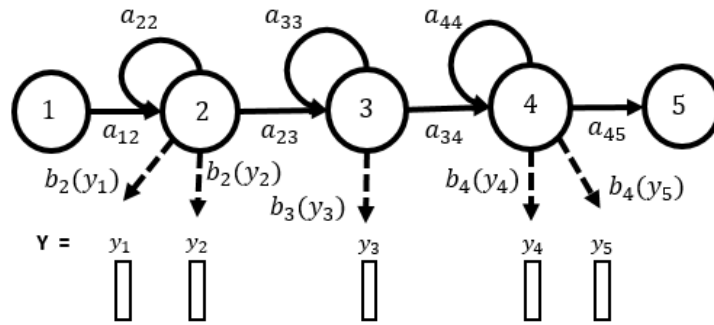
Figure 6. An HMM based phone model with transition probabilities a_{ij} and observation distributions b_j ( )

$$P(Y|w) = \sum_Q p(Y|Q)P(Q|w) , \tag{5}$$

where Q represents a sequence of valid pronunciations. If the assumption of a single multivariate Gaussian is made for the output distribution, then $b_j(y)$ would be given by Eq. 6:

$$b_j(y) = N(y; \mu^{(j)}, \Sigma^{(j)}) , \tag{6}$$

where $\mu^{(j)}$ is the mean of state ω_j and $\Sigma^{(j)}$ represents its covariance. In this sense, the acoustic likelihood is described in Eq. 7:

$$p(Y|Q) = \sum_\theta p(\theta, Y|Q) = a_{\theta_0\theta_1} \prod_{t=1}^T b_{\theta_t}(y_t) a_{\theta_t\theta_{t+1}} , \tag{7}$$

where $\theta = \theta_0, \dots, \theta_{T+1}$ represents a state sequence through the model (Gales & Young, 2007).

The parameters for the acoustic model are commonly estimated through the forward-backward algorithm as explained in Baum et al. (1970). In general, this approach, which is an application of the Expectation Maximization (EM) algorithm, updates the weights of the system to better explain the observed training sequences.

In problems like speech recognition, or electroencephalography in this case, the utilization of a single Gaussian distribution is not necessarily accurate. Variations such as speaker identity, accent, gender and others make this assumption rarely possible in practice (Gales & Young, 2007). To overcome this issue, several systems have successfully implemented mixtures of Gaussians (Young, 1996), which are able to properly, and more accurately, model multi-modal data. If Gaussian mixture models are implemented, then the value for $b_j(y)$ is given by Eq. 8:

$$b_j(y) = \sum_{m=1}^M c_{jm} N(y; \mu^{(jm)}, \Sigma^{(jm)}) , \tag{8}$$

where c_{jm} represents the prior probability for mixture component m of state ω_j. The number of Gaussian mixtures is often optimized experimentally (Gales & Young, 2007).

2.3   Deep Learning Applied to Speech Recognition

Speech recognition was one of the first fields in which Neural Networks (NNs) were applied. It was not until recently, however, that the industry saw a compelling argument for replacing GMM-HMM technology, which could achieve comparable performance levels with less data and computational power (Goodfellow et al., 2017). An example of systems that achieved comparable performance before the 2000s is given by Robinson and Fallside's Recurrent Error Propagation Network (Robinson & Fallside, 1991). The performance on TIMIT for this system was only slightly better than that of the best HMM systems available at the time. Classic GMM-HMM systems, which work by modeling the temporal association between acoustic features and phonemes with HMMs and frequency domain variations of each phoneme with GMMs, dominated the field until the late 2000s, when larger and deeper models were enabled by the availability of more advanced data and computational resources (Goodfellow et al., 2017).

New deep learning approaches based on the training of undirected probabilistic models, better known as Restricted Boltzmann Machines (RBMs), allowed for the effective replacement of GMMs in state-of-the-art speech recognition systems (Goodfellow et al., 2017). Unsupervised pre-training was used to build deep feed-forward networks, whose layers were essentially stacked trained RBMs. These Deep Belief Network (DBN) acoustic models, which use several layers of nonlinear features and generative model pre-training, replaced GMMs and showed a significant improvement on standard speech recognition tasks. For example, the phoneme error rate on TIMIT was decreased from 26% to 20.7% (Mohamed et al., 2012).

Contemporary speech recognition systems have shifted towards techniques that involve the use of Convolutional Neural Networks (CNNs), which treat the input spectrogram as an image, instead of a long vector (Sainath et al., 2013). Another common end-to-end deep learning system

is based on variations of Long Short-Term Recurrent Neural Networks (LSTM RNNs) (Graves et al., 2013; Pascanu et al., 2013; Chung et al., 2014), which usually employ state variables from several layers at each time step, generating two types of depth: a regular depth due to a stack of layers and additional depth due to time unfolding (Goodfellow et al., 2017).

In the following section, Convolutional Neural Networks (CNNs), a type of feed-forward neural network, are discussed. The evolution of these networks for the decoding of sequential data, such as speech, is emphasized, since similar systems are implemented for the decoding of EEGs in this thesis.

2.4   Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a variant of a standard neural networks that, instead of having fully connected hidden layers, presents a network structure that alternates convolution and pooling layers. These networks are characterized by the use of convolution instead of general matrix multiplication in at least one of their layers (Goodfellow et al., 2017). The purpose of this section is to explain the general theory behind CNNs and highlight the role these networks have played in the development of modern deep learning-based speech recognition systems.

CNNs, unlike other traditional feedforward networks, do not map the whole input space with different weight vectors to each unit in the first hidden layer. Instead, they represent the presence or absence of a feature in a local subspace using a filter, which is replicated across the whole input space producing a feature map (Mohamed, 2014). These feature maps can commonly be described by Eq. 9 for an input vector x:

$$h_{ij}^k = f((W^k * x)_{ij} + b_k) \,. \tag{9}$$

In this case, h^k represents the kth feature map at a given layer and W^k and b_k represent the weights and biases respectively. These variables are arguments to the function f which is commonly a non-linearity such as a $tanh, rectifier$ or $sigmoid$ function. As the name of the networks suggests, the input is convolved with the filter (denoted by '*' in Eq. 9).

Convolution leverages three useful aspects that could potentially improve performance and efficiency: sparse interactions, parameter sharing, and equivariant representations. Traditional feedforward neural networks do matrix multiplication of a matrix of parameters with a separate parameter describing the interaction between each input and output unit (Goodfellow et al., 2017). In other words, every input interacts with every output. With CNNs, however, the kernels (also referred to as filters), are usually smaller than the input, allowing for the detection of meaningful features in each portion (e.g.: edges of an image) of the input. This property, often called sparse interaction or sparse connectivity, introduces efficiency in the process, since fewer parameters need to be stored and fewer operations need to be performed (Goodfellow et al., 2017).

CNN runs a small window over the input so that the weights of the network can learn from various features of the input data regardless of the absolute location of the feature. In other words, all of the hidden units that are forming a feature map, share the same filter (Mohamed, 2014). Traditional neural networks differ from CNNs in that each element of their weight matrix is used only once for the computation of the output layer. CNNs, on the other hand, have tied weights, because the value of a weight applied to one input is the same as a weight applied somewhere else in the input, reducing the storage requirements for the model (Goodfellow et al., 2017).

The shared parameters introduce another property called equivariance to translation. Equivariance essentially implies that if the input changes, the output changes in the same way. This can be explained in the context of time series data. Convolution produces a timeline that

demonstrates when different features appear in the input. If an event appears later in time in the input, the exact same representation will show in the output (Goodfellow et al., 2017). The combination of these properties allows the system to model minor translations in the input.

CNNs systems often alternate between convolutional and pooling layers. The purpose of the pooling layer is to subsample the feature map produced by the convolutional layer. This generates a lower resolution feature map that is robust to small deviations in feature locations (Mohamed, 2014). Commonly used pooling operators include the calculation of an average or the maximum value over a pooling window.

CNNs have been previously applied to acoustic modeling. For example, CNNs have been used to learn more stable acoustic features for tasks such as gender, speaker and phone classification (Hau & Chen, 2011; Lee et al., 2009). In these projects, convolution was applied to windows of acoustic frames that overlapped in time in order to learn stable acoustic features that would be useful for classification tasks. Abdel-Hamid et al. (2014) extended this concept by introducing a hybrid CNN-HMM framework that successfully used a softmax output layer on top of a CNN in order to compute the posterior probabilities for all HMM states.

In general, the properties of CNNs offer some advantages for automatic speech recognition. First, the locality in the units of the convolutional layers increases robustness against non-white noise (Abdel-Hamid et al., 2014). This property allows good features to be computed locally from higher signal to noise frequency bands. This represents a potentially beneficial feature for the decoding of EEGs, where some frequency bands present exclusive types of noise, such as muscle artifacts in the beta and gamma ranges (Muthukumaraswamy, 2013).

Another advantage offered by CNN for the speech recognition task relates to weight sharing. The weights are learned from different parts of the spectrum, creating a more robust model

and a model that is less prone to overfitting. In addition, since pooling is performed, feature values computed at different locations are pooled together and represented by a single value (depending on the pooling operator used), which minimizes the differences between input patterns when there are slight frequency shifts. In speech, these frequency shifts can be attributed to factors such as the differences in vocal tract length for different speakers, while with EEGs, these shifts can be a result of the subject's age (Ebersole & Pedley, 2014).

2.5   Multilayer Perceptron (MLP)

It is clear that many of the benefits of CNN rely on the fact that the convolutional layers are only connected to a local section of the input. Getting a probability distribution for each labeled class in the data while using a CNN, therefore, is usually done through the implementation of a layer that has full connectivity to all of the activations in previous layers. This fully-connected layer has been successfully implemented in the form of a traditional Multilayer Perceptron (MLP) in systems such as those developed by Krizhevsky et al. (2012) and Girshick (2016).

As classifiers, MLPs can be defined as feedforward neural networks that map an input x to a category y through the approximation of a function f^* and through learning the parameters θ in $f^*(x, \theta)$ (Goodfellow et al., 2017). MLPs can be used as classifiers in which the input is first projected into a linearly separable space, as shown in Eq. 10, through the implementation of a learned non-linear transformation ϕ that relates to the function f^*:

$$y = f(x; \theta, w) = \phi(x; \theta)^T w \,, \tag{10}$$

where w are the parameters that map x using $\phi(x)$ to the output. Similar to the CNN activation functions, typical choices for f include the $sigmoid$ and $tanh$ nonlinearities (Goodfellow, 2017).

Table 1. Summary of word error rates for a subspace GMM and a Deep NN

Corpus	Training Speech	SGMM WER	DNN WER
BABEL Pashto	10 hours	69.2%	67.6%
BABEL Pashto	80 hours	50.2%	42.3%
Fisher English	2000 hours	15.4%	10.3%

2.6 Performance Comparison of GMMs-HMMs and Deep Neural Networks (DNNs)

Over the last decade, advances in computer hardware, machine learning, and deep learning algorithms have facilitated faster and more accurate training of Deep Neural Networks (DNN) (Hinton et al., 2012). As described before, this technology has made a series of breakthroughs in the areas of speech and image processing recently, outperforming systems based on approaches such as HMM and GMM-HMM. The performance gap, however, gets smaller as the amount of training data decreases. This observation is evident from the results that have been obtained with the Kaldi Speech Recognition Toolkit (Povey et al., 2011) on the Intelligence Advanced Research Projects Activity (IARPA) provided database, BABEL, and the Fisher English Corpus. These results are summarized in Table 1.

The results in Table 1 show that, indeed, DNNs are capable of achieving significant improvements in the performance of a speech recognition system. However, the difference is not as significant as when the number of training observations is not large enough. Hence, a significant part of this thesis is the development of a big data resource to support training of deep learning systems.

CHAPTER 3

DATA PREPARATION

This study utilized a subset of the Temple University Hospital EEG (TUH EEG) Corpus, which represents the largest publicly available database of clinical EEGs (Harati et al., 2014). The database is currently comprised of more than 30,000 records from over 18,000 unique patients. Given the nature of this study, it is important to emphasize that 75% of the records present in TUH EEG are classified as abnormal in the EEG reports provided with the data (and part of the patient's medical record). In this chapter, the process of constructing the data sets used to support our experimentation is described.

In this chapter were introduce two important subsets of the data, referred to as a short set and a full set. The short set was used for a preliminary investigation of the problem. The machine learning systems described in this thesis, particularly the HMM and deep learning systems, are computationally very expensive. Short sets are very useful for rapid turnarounds on experiments, allowing quick evaluation of many design tradeoffs. The full dataset, on the other hand, was used to train and test the final models. Because of the large number of parameters required by the deep learning models, these models were only evaluated on the full dataset. A short set must be designed carefully so that it yields performance that is representative of the larger set.

3.1   The Short Dataset

For the purposes of the study, a demographically balanced short subset of the TUH EEG database was selected. This database was specifically used for the pilot studies and the selection of an appropriate model for the baseline. The age and gender of the patients were considered for the selection of the data. Because pediatric EEGs are very different in nature from adult EEGs (Ebersole & Pedley, 2014), the majority of the records utilized were obtained from patients that

were older than 20 years old. Figure 7 provides histograms of ages for the training and evaluation

sets respectively. It is possible to see that, excluding a small number of outliers in the datasets, all

of the patients are in the age range of 20-90, with a mean of 46.6 years and a standard deviation of

14.7 years. The genders of the patients, as it can be also seen in Figure 7, were also kept balanced.

The selected data was divided into two sets: a training set, which contained 80 abnormal

and 82 normal EEGs, and an evaluation set, which contained 55 abnormal and 51 normal EEGs.

From these recordings, only one channel was considered for the final analysis. This channel was

selected through experimentation using a process that is described in Section 4.2.

3.2   The Full Dataset

The short dataset was useful for the execution of some pilot experiments, the selection of

appropriate models, and the explorations of relevant channels. However, the training set was not

considered to be big enough for the training of more complex models, such as deep learning

systems. In addition, the utilization of a larger dataset would ensure a more heterogeneous set and,

Figure 7. Distribution of the patients' ages and genders for the short dataset. *a)* Gender distribution
of the training dataset; *b)* Age distribution for the training dataset; *c)* Gender distribution for the
evaluation dataset and *d)* Age distribution for the evaluation dataset.

therefore, a good source of data for the training of more robust models that would operate under a multitude of conditions.

This database was developed by screening all medical reports using natural language processing (NLP) of all sessions recorded with an Averaged Reference (AR) electrode configuration (Lopez et al., 2016). The AR electrode accounts for about 45% of the data in the overall corpus. After all the records were classified as either normal or abnormal from the information provided in the medical report, a team of students manually reviewed each signal and its respective report to ensure the initial class assigned through the NLP step was correct. Demographic information for the full database is provided in Figure 8, while Table 2 and

Table *3* present some descriptive statistics for the evaluation and training partitions respectively.

The TUH EEG Corpus is divided in EEG sessions, which contain several pruned EDF files. For the purposes of the normal/abnormal EEG discrimination, only one EDF file was selected for each session. Each file selected from a session was chosen by considering the length of the files
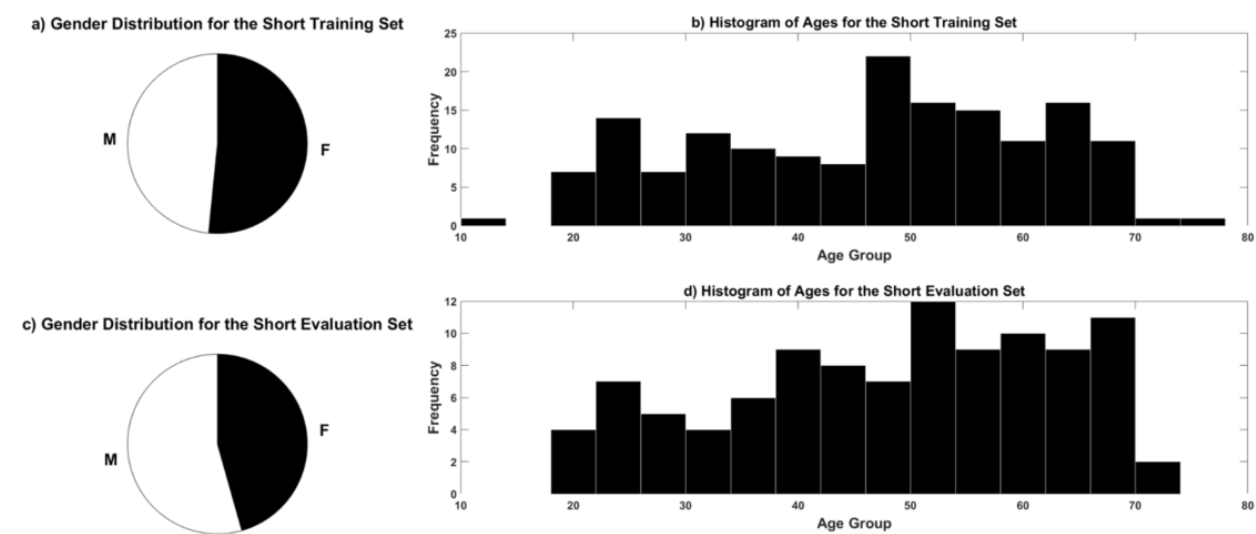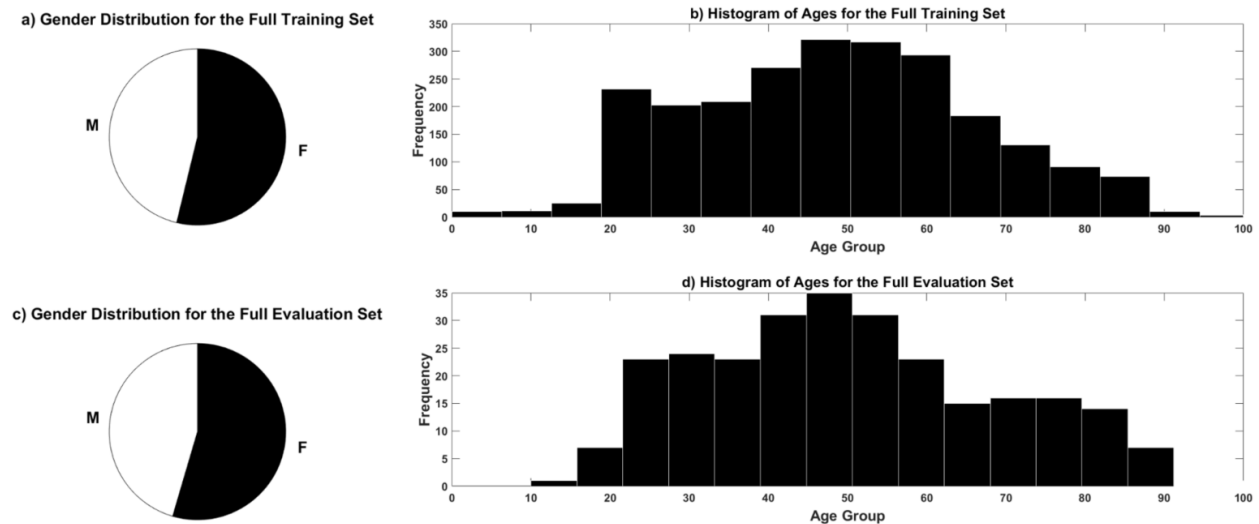

Figure 8. Distribution of the patients' ages and genders for the full dataset. *a)* Gender distribution of the training dataset; *b)* Age distribution for the training dataset; *c)* Gender distribution for the evaluation dataset and *d)* Age distribution for the evaluation dataset.

Table 2. File statistics for the full evaluation set.

Description	Files		Patients		hours
Evaluation					
Abnormal	130	46.4%	105	41.5%	48.9
Normal	150	53.6%	148	58.5%	55.4
Total	280	100.0%	253	100.0%	104.4

Table 3. File statistics for the full training set.

Description	Files		Patients		Hours
Training					
Abnormal	1398	50.2%	899	42.1%	546.4
Normal	1387	49.8%	1239	58.0%	518.3
Total	2785	100.0%	2138	100.0%	1064.7

(all the files of the database used are longer than 15 minutes) and/or the presence of relevant activity.

3.3 Features

Feature extraction was performed on the EEG data in a preprocessing step. The feature extraction approach followed techniques that are similar to techniques based on Mel Frequency Cepstral Coefficients (MFCCs) that have been used for speech recognition (Picone, 1990). MFCCs are normally calculated through the computation of a high resolution Fast Fourier Transform and down-sampling the results with an oversampling approach that uses overlapping bandpass filters. The results obtained from this process are then transformed to the cepstral domain through a cosine transform (Huang et al., 2001). When extracting the cepstral coefficients from EEG signals, a very similar approach is followed, with the exception that the filter used to convert the spectrum to filterbank amplitudes are linearly spaced. Only the first eight cepstral coefficients were retained. The energy of the signal was calculated in the frequency domain and used to replace the 0th-order cepstral coefficient. The calculation of the frequency energy is given by:

$$E_f = \log \left(\sum_k^{N-1} |X(k)|^2 \right) . \tag{11}$$

It is important to note that the frame and window duration for this portion of the feature extraction is 0.1 seconds and 0.2 seconds respectively (Harati et al., 2015).

The extraction of the frequency energy and the cepstral coefficients is followed by the calculation of another type of energy: the differential energy (E_d). Differential energy is a feature derived from the features that have been described to this point, and it is given by the difference between the largest and the smallest sample in a 0.9 seconds window. This feature is computed using the following approach:

$$E_d = \max_m\big(E_f(m)\big) - \min_m\big(E_f(m)\big).$$
(12)

Here, *M* represents the number of frames. Figure 9 illustrates the feature extraction process.

The first and second derivatives (differential and acceleration coefficients) of the base features, often referred to as absolute features, are also computed and concatenated to the feature vector. These features represent the trajectory of the base features and are calculated using a regression analysis approach (Huang et al., 2001):
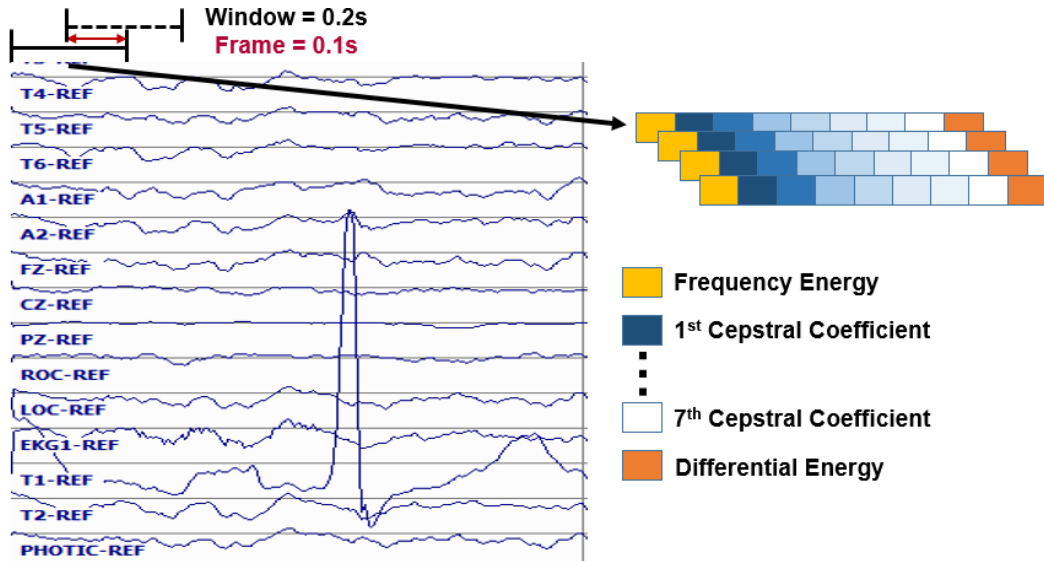
Figure 9. Illustration of the base feature extraction process.

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2 \sum_{n=1}^{N} n^2} . \tag{13}$$

A delta coefficient, d_t, represents a derivative calculated for frame t in terms of the static coefficients c_{t+n} to c_{t-n}. Similar to the calculation of the differential energy feature, E_d, the window N used in the calculation of the first and second derivatives is set to 0.9 seconds in this study (Harati et al., 2015).

In summary, the features that are extracted from the EEG signals are the frequency energy (1 feature), 7 cepstral coefficients (7 features), a differential energy term (1 feature), and the first and second derivatives (17 features) of the base features. Note that the second derivative is not extracted for the differential energy base feature, since this feature is redundant and did not show a positive impact in performance (Harati et al., 2015). This decreases the overall dimension of the feature vector to 26.

3.4   Dimensionality Reduction

In this research, we attempted to achieve high performance using only one channel of the signal and using only the first T seconds of the signal. This experimental paradigm was based on the fact that neurologists are reportedly able to distinguish a normal recording from a normal one by examining first few seconds of the files. To establish a baseline for the project, the first 60 seconds of the signal were considered, and the features across all the frames corresponding to this interval were stacked together, forming a 15,600 (600×26) feature vector (Lopez et al., 2016).

This is obviously an excessively high dimensionality. If all channels were used, the dimensionality would increase by another factor of 20. Hence, it was decided that a single channel would be used to control dimensionality without compromising the performance of the system. Even with this reduction, it was still necessary to do dimensionality reduction of this vector using

Principal Components Analysis (PCA) (Jolliffe, 2002). These experiments are described in the next chapter.

CHAPTER 4

EXPERIMENTAL RESULTS

In this chapter, we present the results of the experiments performed for this thesis. First, we present the pilot experiments, which involved the classification of the normal and abnormal EEGs with the kNN and the RF algorithms. Next, we demonstrate how we optimized and evaluated our HMM baseline system. Finally, we present the improvements that we obtained through the integration of deep learning technology. In each section, we explain the system that was evaluated and then report on the results achieved with that system.

4.1   Baseline Results: k-Nearest Neighbor and Random Forest

In order to establish baseline performance for this task, we evaluated two standard machine learning algorithms: k-Nearest Neighbors (kNN) and Random Forest Ensemble Learning (RF). We selected these systems because they have provided extremely stable performance over a wide range of machine learning tasks. Our initial experiments were conducted only with the short dataset. We sequentially optimized key parameters of each component, and then jointly optimized the resulting systems.

The first set of experiments involved exploring dimensionality reduction in PCA. Prior to doing this we needed to optimize one key parameter for kNN and RF. For kNN, the number of nearest neighbors (k) was varied from 1 to 100. For RF, the number of trees, N_t, was varied from 1 to 100. Once the parameters were properly optimized, we studied the optimal PCA dimension and which single channel was best for classification. The optimized systems were tested for each one of the 22 channels in the transverse central parietal (TCP) montage (ACNS, 2006), which accentuates spike activity. The results of these experiments helped us establish an initial baseline for the classification of normal and abnormal adult EEGs.

Figure 10 demonstrates that performance for RF seems to saturate for $N_t > 50$. We selected $N_t = 50$ as a compromise between training time and performance (Lopez et al., 2015). Figure 11 demonstrates performance as a function of the number of nearest neighbors (k). Though performance fluctuates considerably, $k = 20$ provides reasonable performance. Figure 12 demonstrates performance as a function of the PCA dimension. The previously selected number of trees $N_t = 50$ was used for the RF implementation, while a value of $k = 1$ was used for kNN (Lopez et al., 2015). For kNN, the performance is not heavily impacted by PCA values larger than 20, while for RF, the error decreases slightly for large values of the PCA dimension. A value of 86 was selected for the input dimension.

Finally, we used these optimized models to explore channel selection. We used 22 channels of the TCP montage individually for the classification. In this way, it was possible to understand which channels (and what regions of the scalp) contribute the most for the
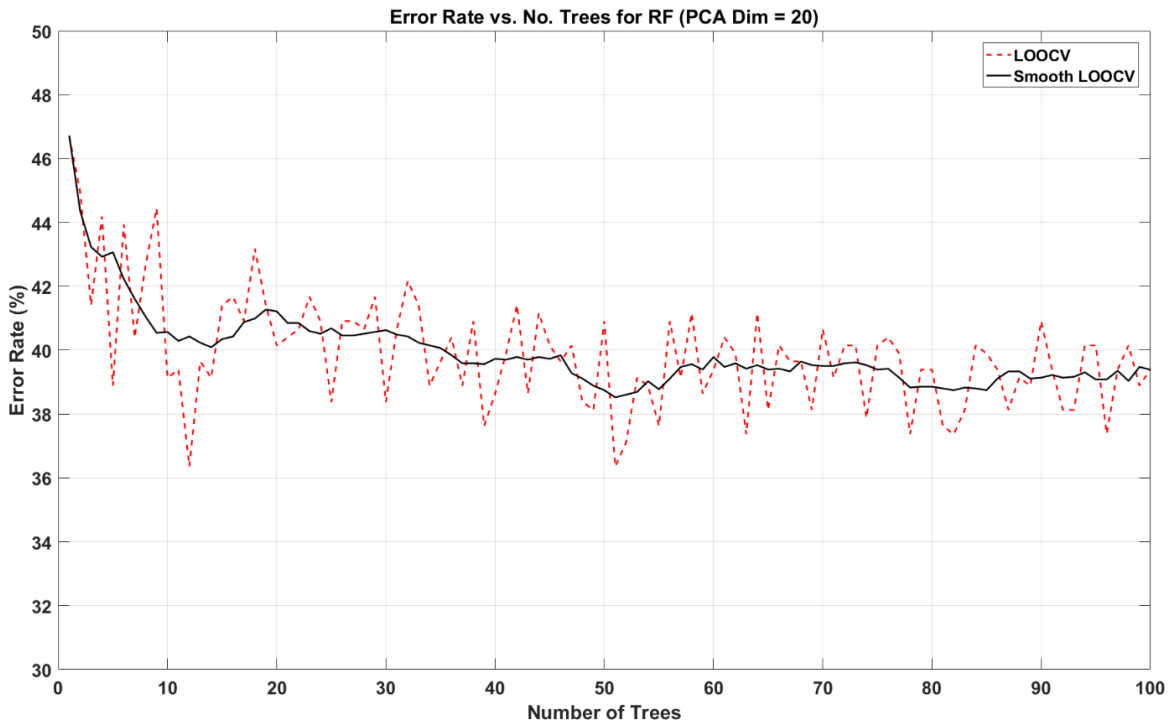
Figure 10. Normal/abnormal classification error rate as a function of number of (trees N_t)

identification of abnormal EEGs. Figure 13 shows the most relevant results from this study. This analysis was performed using the optimized kNN system ($k = 20$), since it presented less variance than the RF implementation (Lopez et al., 2015).

The performance reported in Figure 13 is based on the channel that performed worse (F4-C4) and the channel that performed best (T5-O1). These results are consistent with the way in which neurologists interpret EEGs, which usually involves the identification of abnormalities (slowing, lack of reactivity) in the posterior dominant rhythm (PDR), present in the posterior regions of the scalp (posterior, occipital-temporal channels in the TCP montage). In this sense, since T5-O1 is a temporal-occipital channel, it shows the PDR reactivity (changes in the signal's properties when patients open and close their eyes) more markedly and, as the preliminary results confirm, better represents the instances of the file that indicate some EEG abnormality.

To summarize, the performance of the optimized version of both systems for varying input
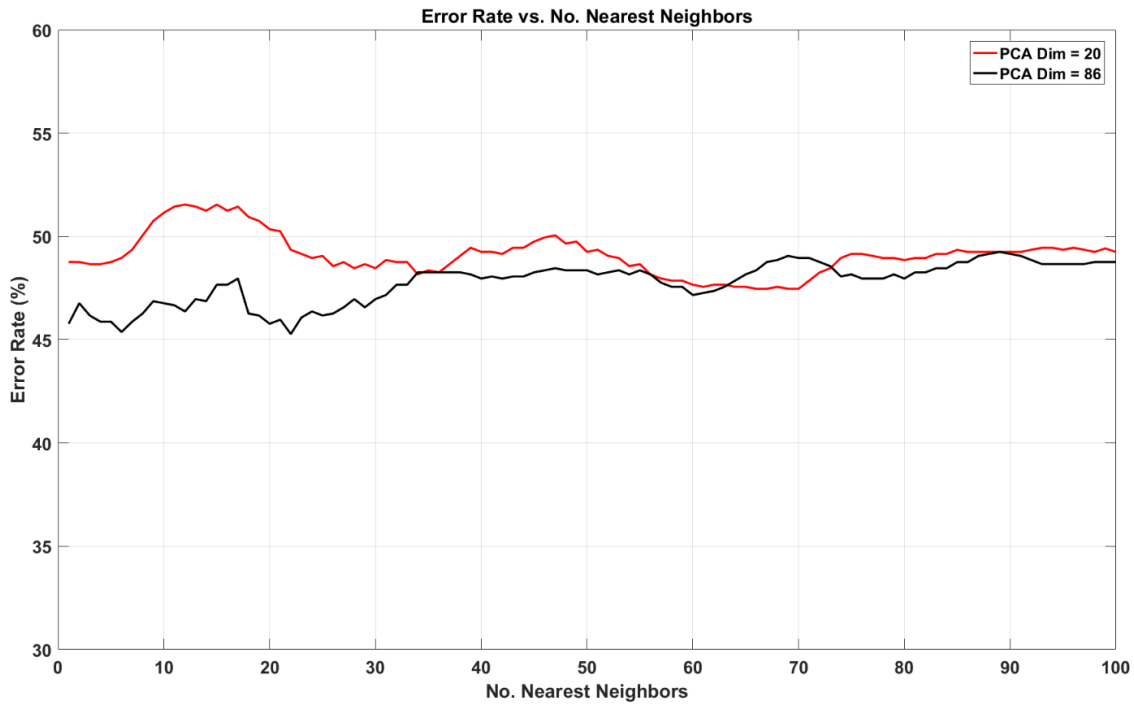
Figure 11. Error rate as a function of the number of neighbors k for PCA dimension of 20 and 86

28
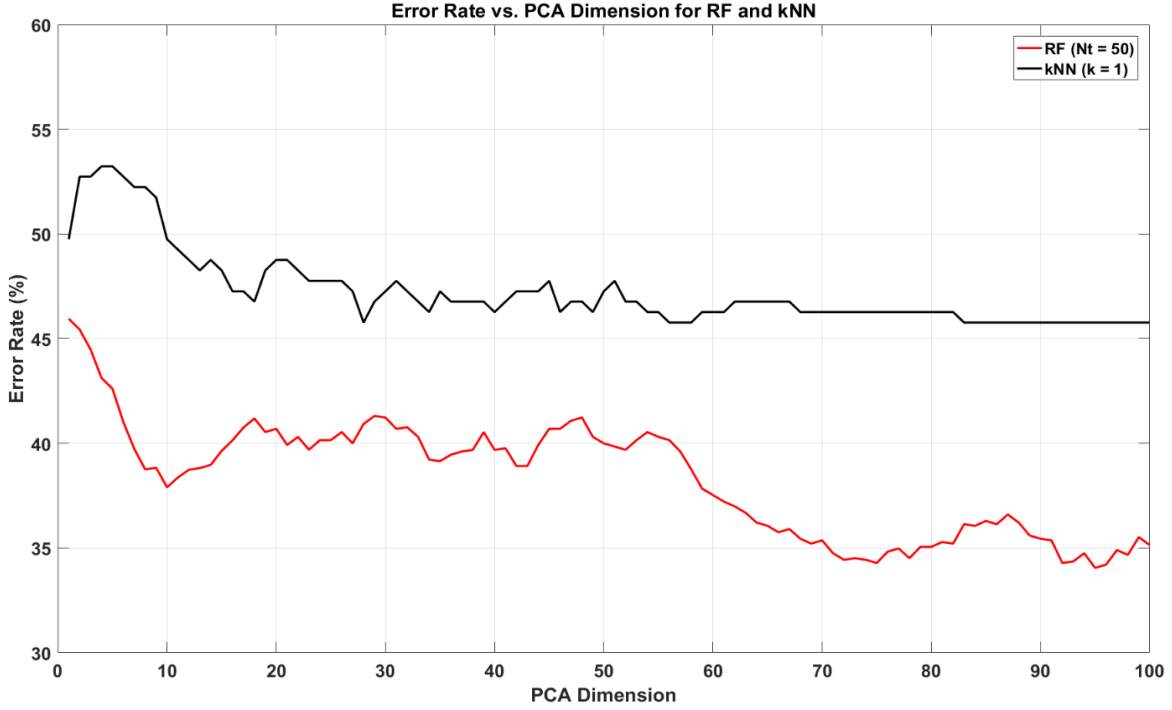
Figure 12. Error rate as a function of PCA dimension

feature vectors dimension is presented in Figure 12 and summarized in Table 4. The RF algorithm

showed considerably higher variance than the kNN system. A confusion matrix for kNN is also

shown in Table 5. These results show that the dominant error for this system occurs when a normal

EEG is classified as abnormal. This can be attributed to the presence of benign variants, which are

electrographic patterns that resemble abnormalities but do not trigger an abnormal classification.

In the next section, we introduce a series of HMM systems that we implemented in order to better

Table 4. Comparison of the performance obtained with the two baseline systems

No.	System Description	Error
1	kNN ($k = 20$)	41.8%
2	RF ($N_t = 50$)	31.7%

Table 5. Confusion matrix for the kNN system

Ref/Hyp	Normal	Abnormal
Normal	50.5%	49.5%
Abnormal	34.0%	66.0%
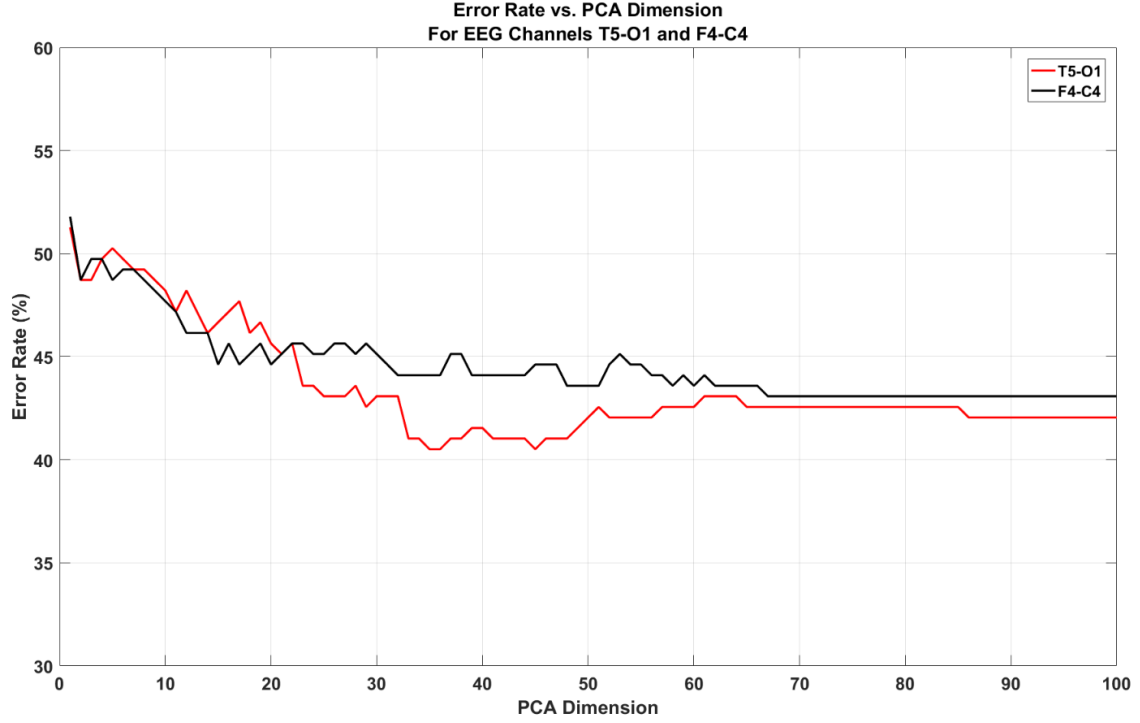
Figure 13. Classification error rate (for kNN) for a fronto-central (F4-C4) and a temporal-occipital (T5-O1) channel

exploit the sequential nature of the signal. We hypothesize that more advanced temporal modeling would decrease the negative impact introduced by benign variants.

4.2   GMM-HMM Systems

Our next attempt to improve performance consisted of designing and evaluating three fundamental variations of HMM systems: (1) a system that sequentially analyzes windows of data for the first t seconds in a signal and outputs a final decision; (2) a system that decodes each one second epoch, outputs a classification decision for every epoch, and postprocesses these decisions with a majority vote; and (3) a hybrid HMM-SdA system that utilizes an HMM to make a classification decision for every epoch and then post-processes the posterior probabilities with a deep learning system. It is important to note that all these systems operated with the features extracted as explained in Section 3.3 and only with information from one channel – T5-O1. We

then conducted a comparative analysis with channels representative of different scalp regions.

Several experiments were conducted in order to optimize the HMM system such as optimizing the number of Gaussian mixtures. To avoid the problem of overfitting, the short dataset and system no. 1 were used for the optimization of the HMM portion of each system. The first step for these experiments was to find the optimal number of Gaussian mixtures and states for the HMM by running classification experiments with the full set of features and the first 10 minutes for each file. Once the system's parameters were properly optimized, we used the models to find the optimal amount of input time for the signal by varying the input time from 5 minutes to 25 minutes in steps of 5 minutes.

The optimized HMM system based on system no. 1 was also implemented using feature vectors reduced in dimension by PCA in order to fairly compare this system with the pilot kNN and RF studies. Once system no. 1 was fully optimized, we evaluated performance for several electrode locations as shown in Figure 14. We selected channels that represented at least one
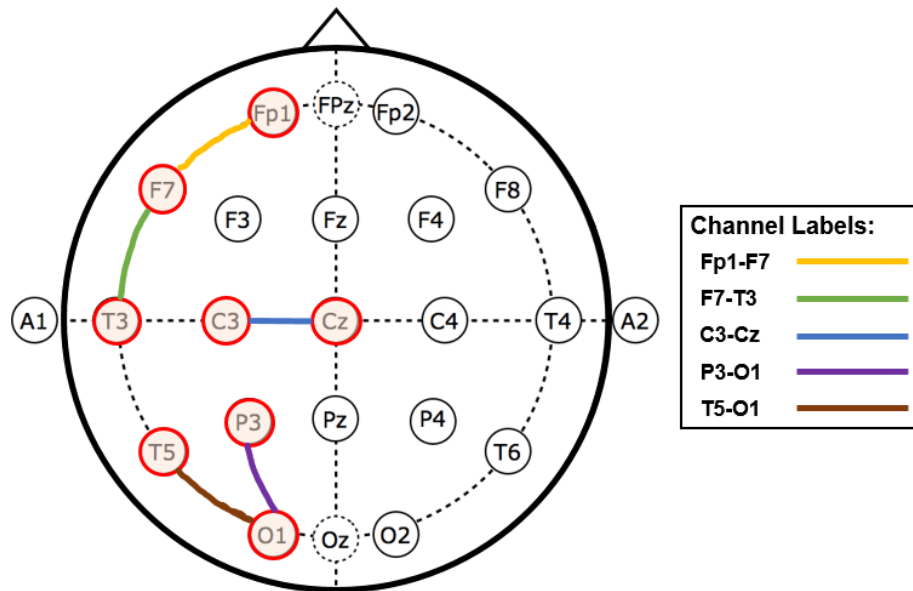
Figure 14. Location of studied channels in the 10-20 standard system of electrode placement for the TCP montage

general area of the scalp (frontal, parietal, temporal or occipital). Since our analysis is mostly focused on the background EEG, which tends to be symmetric, only the left side of the scalp was analyzed.

The channel test with the HMM system was conducted in order to verify whether the channel optimization could be generalized. Since the optimization process for the GMM-HMM system was conducted with the short database, the best system was then trained and evaluated with the full dataset in order to validate the design on previously unseen data. Systems no. 2 and 3 decode one-second epochs at a time. The output generated by this system was then post-processed with selected methods in order to obtain a final decision output. Two different post-processing methods were used for this system: a majority vote simple baseline (2) and a Stacked denoising Autoencoder (SdA). We refer to the latter system as GMM-HMM-SdA.

Stacked denoising Autoencoders (SdAs) are formed by stacking denoising autoencoders to form a deep network. The selection of SdAs for the post-processing technique was based on the fact that SdAs are designed to be robust to partial destruction (or missing modalities) (Vincent & Larochelle, 2010). They are a good option for signals such as EEGs, which can often exhibit noisy or non-ideal conditions. Windows of the input data along with the posterior probabilities decoded with the GMM-HMM step for each epoch were used as an input to the SdA post-processing system. This system output a normal/abnormal decision for every one-second epoch based on a longer window duration than in the GMM-HMM pass, additionally giving the system a higher temporal context for each decision. The results obtained through the GMM-HMM-SdA method were then compared to a simple baseline that was based on majority vote. Essentially, the normal/abnormal decision for the file was made by selecting the higher number of occurrences for each class in a given file.

The optimization of GMM-HMM system no. 1 for this classification problem involves the selection of parameters such as the number of Gaussian mixtures and the number of HMM states. The first 10 minutes of data (features) for the T5-O1 channel were used as an input to the system. Table 6 gives a summary of the open-loop results that were obtained through the evaluation of several system parameters. The closed loop performance for the best system (#GMM = 3, #HMM = 3) achieved an error rate of 13.6%.

To understand how much signal information would work better for the identification of abnormal EEGs, the optimized system was used to process different input lengths. Table 7 shows this analysis, and reveals that the best performance can be obtained for an input time of 10 minutes. The length of most the recordings in the dataset are less than 25 minutes, so the performance saturates for durations longer than 25 minutes.

Thus far the results that were presented were calculated with data from the T5-O1 channel, which was found to be optimal for the baseline systems. To make sure the channel selection was still optimal, we explore performance as a function of channel in Table 8. We observe that the channel that performed best for the GMM-HMM system is the same that was discovered through

Table 6. GMM-HMM open-loop error rates for various HMM parameters

# Gaussian Mixtures	# HMM States	Correct Detection (%)
1	1	30.2%
1	2	34.9%
1	3	34.9%
2	1	23.6%
2	2	19.8%
2	3	22.6%
3	1	23.6%
3	2	17.9%
3	**3**	**17.0%**
4	1	17.9%
4	2	35.9%
4	3	22.6%

Table 7. Correct detection rate for different signal input lengths

Input (min)	#Gaussians/#HMM States	Correct Detection (%)
5	3/3	19.8%
10	**3/3**	**17.0%**
15	3/3	19.8%
20	3/3	20.8%
25	3/3	23.6%

the baseline system experiments – T5-O1. In this sense, it can be said that this temporal-occipital channel has great relevance in the classification of abnormal EEGs.

The results that have been presented to this point, can be further summarized and compared to the baseline performance. Table 9 shows the results of this comparison. The PCA-HMM experiment used the same exact inputs that were used for the baseline systems (for comparison). It can be seen that the best performance was achieved by both of the HMM systems, with the full feature system having the lowest overall error rate. Table 10 provides a confusion matrix for the best reported system. The false alarm rate, which is defined as the rate of normal EEGs classified as abnormal, of the GMM-HMM system showed an improvement of 27.7% compared to the false alarm rate of the baseline kNN system. The HMM systems did a better job at modelling the signals and discriminating EEG records in which benign variants could potentially confuse the system.

Thus far, a series of HMM-based models have been evaluated on the short database. In order to further test the fit of these models, we evaluated selected models on the full dataset described in Section 3.2. The error rate, as it was expected, increased significantly with the

Table 8. Correct detection rate for different channels

#Gaussians/#HMM States	Channel	Correct Detection (%)
3/3	Fp1-F7	35.8%
3/3	T5-O1	**17.0%**
3/3	F7-T3	23.6%
3/3	C3-Cz	18.9%
3/3	P3-O1	20.8%

Table 9. Summary of the performance for all the evaluated systems

System Description	Error (%)
kNN (k=20)	41.8%
RF (Nt=50)	31.7%
PCA-HMM #GM = 3 #HMM States = 3)	25.6%
GMM-HMM (#GM = 3 #HMM States = 3)	**17.0%**

introduction of new data that contained a variety of conditions. This system yielded an error rate

of 26.1%. The addition of new data significantly increased the error rate that was obtained with

the short dataset, showing that the generated GMM-HMM model is not complex enough to

properly explain the data. In this sense, steps were taken to increase the model complexity with

the full dataset.

First, a hybrid HMM system, denoted system no. 3, was implemented and compared to

system no. 2. This hybrid HMM-SdA system was able to achieve an error rate of 22.9%. As it can

be seen in Table 11, this model improved the detection of normal EEGs by ~9%, while decreasing

the abnormal detection rate by ~3%. In essence, Table 11 indicates that the normal EEG correct

detections present a significant improvement from the baseline, while the classification of

abnormal files shows a higher error rate. Overall, the GMM-HMM-SdA system performed better

than the pure GMM-HMM system, and also surpassed the GMM-HMM system using majority

vote, which yielded an error rate of 24.6%. These observations show that the increased complexity

of the system was beneficial to the classification task, and justify the implementation of more

complex systems.

The performance of the system improved with the more complex models. However, the

results were not significantly better than the pure GMM-HMM system. In order to achieve

Table 10. Confusion matrix for the best GMM-HMM system (Short Dataset)

Ref/Hyp	Normal	Abnormal
Normal	78.2%	21.8%
Abnormal	11.8%	88.2%

Table 11. Confusion matrix for GMM-HMM-(SdA) system

Ref/Hyp	Normal	Abnormal
Normal	90.0%	10.0%
Abnormal	37.7%	62.3%

additional performance improvements, a more complex end-to-end CNN-MLP deep learning system was trained and tested with the full TUH EEG Abnormal database.

4.3   A CNN-MLP System

The advantages that CNNs present for speech recognition, which were outlined in Chapter 2, were the motivation for the development of an end-to-end deep learning system. A 2D CNN was trained and evaluated with the features described before for the full dataset. Figure 15 depicts the architecture for a CNN system designed to better exploit temporal behavior in the signal. The features that correspond to windows of data that last t seconds long serve as inputs to the network. Multiple channels can be input to the system allowing modeling of spatial correlations.

Elements of the design, such as the resolution of the input and the depth of the network were evaluated through experimentation. In addition, the premise that the abnormality of an EEG can be better evaluated better from certain areas of the scalp, rather than a single channel, was investigated. In this case, however, instead of using only one channel for these tests, 4 channels
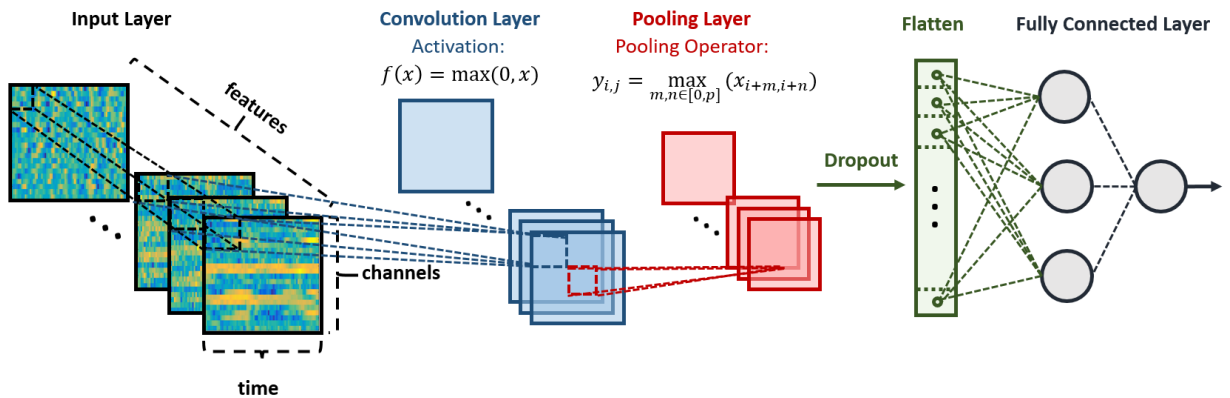
Figure 15. Representation of the input and layers for the CNN System

from each area of the scalp were selected for the evaluation of each system. The information from four areas of the scalp was isolated and utilized individually for the training and evaluation of the system. For simplicity, these areas are called Region I (frontal), Region II (Centro-temporal), Region III (Centro-temporal-parietal), Region IV (Temporal-occipital). Figure 16 shows an illustration of the scalp regions that were individually tested with the CNN system.

To justify the use of deep architectures for this problem, it was necessary to compare performance to shallower versions of the architecture. To accomplish this, the depth of the CNN-MLP network was analyzed. For simplicity, a convolutional layer (L_x), where x is the layer depth was defined as follows:

1. 2D Convolutional layer with ReLU activation

2. 2D Convolutional layer with ReLU activation

3. Maxpooling (2, 2)

4. Dropout (25%)

After the last convolutional layer, every system was fully connected to a fully connected MLP layer. The fully connected layer F can be defined as follows:

Figure 16. Diagram that shows the regions of the scalp (Regions I-IV) that were individually processed with the CNN end-to-end deep learning system.

1. Flattening Layer

2. Fully Connected MLP with ReLU activation

3. Dropout(50%)
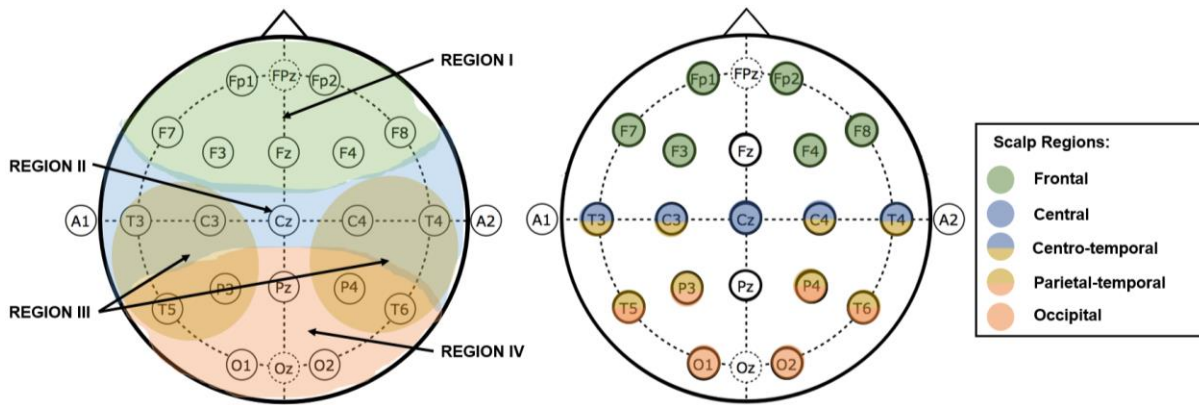
4. Fully Connected MLP with Softmax activation

The activation functions used through the systems, ReLU and Softmax, are defined in Eq. 14 and Eq. 15 respectively (Goodfellow et al., 2017):

$$f(x) = \max(0, x), \tag{14}$$

$$softmax(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}. \tag{15}$$

The number of stacked convolutional layers was varied in order to find the optimal depth for the classification system. A fully connected layer with $Softmax$ activation was connected to the last convolutional layer in order to output a class probability.

Before the main experiments were conducted, the closed loop performance of the system was studied as a function of training iterations. In order to observe the overfitting of the objective function, the dropout regularization layers for the CNN were deactivated for the experiment. In this way, the function would not be modified towards good generalization, and the overfitting of the system would be ideally observed (Goodfellow, 2017). The behavior of the system trained with two different optimizers, basic Stochastic Gradient Descend (SGD) (Bottou, 2010) and Adaptive moment estimation (*Adam*) (Kingma & Ba, 2015), was analyzed as a function of the number of training epochs.

The behavior of the systems as the training iterations increased are shown in Figure 17. The initial iterations show the trend that would be expected in this type of analysis. As the number of training iterations increase, the training error decreases, and the evaluation error decreases as well, until the system starts overfitting (8th iteration). After the 8th iteration, however, the error rate
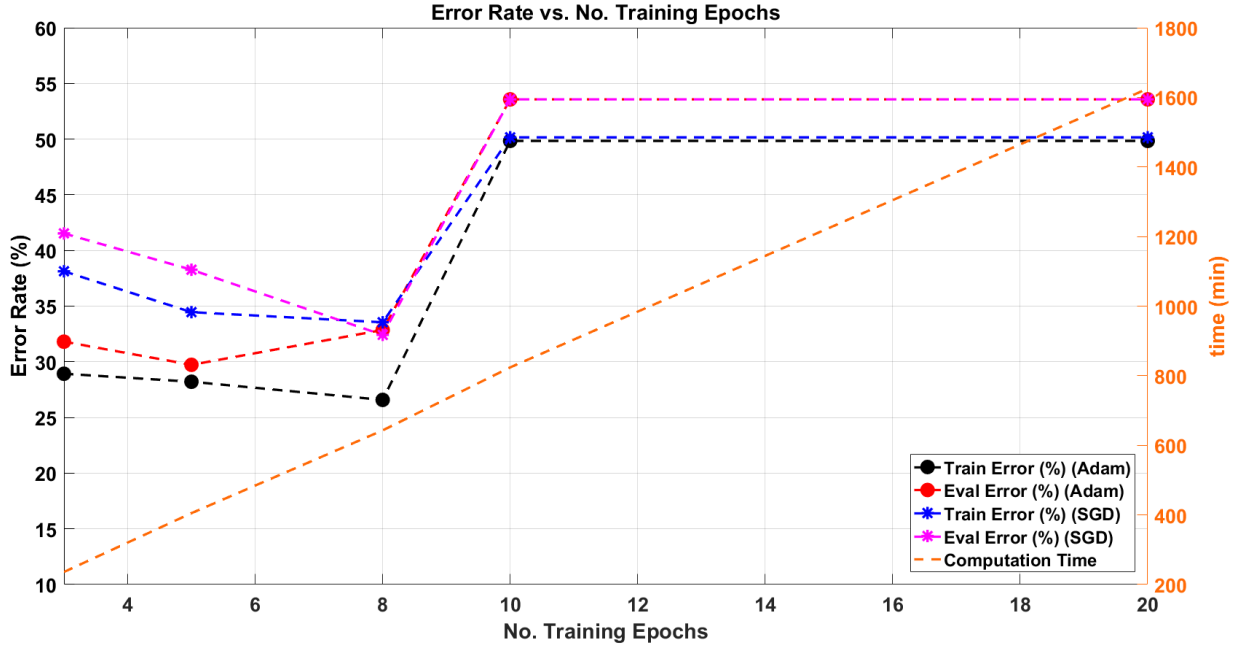
Figure 17. Error rate as a function of the number of training epochs for SGD and *Adam* optimizer. The figure additionally shows the training time as a function of the number of training epochs.

of the system dramatically increases and the performance saturates for both optimizers. The increase in the error rate and performance saturation in the later iterations is not expected behavior, and will require further investigation.

The error rates that were obtained through the different network depths are reported in Table 13. From this summary, it is possible to see that the three-layer representation yielded the best results for the abnormal EEG classification task. It can also be seen that the shallower network produced the worst results out of all the systems. With these results, it is possible to justify the use of a deep learning architecture for the problem.

The resolution of the input was also investigated. The window length was varied in several optimization experiments. A system with three convolutional layers was implemented in order to conduct the window duration analysis. Table 12 summarizes the results for the window duration experiments, and shows that the best results are obtained with a window duration of 7 seconds. In

Table 12. Window duration analysis for the input of the CNN

Window Duration	# Convolutional Layers	Error (%)
3 seconds	3	55.5%
5 seconds	3	46.6%
7 seconds	**3**	**21.2%**
9 seconds	3	26.2%

practice, the analysis of EEGs is conducted through the observation of 10-second windows of data

and the analysis of the presence or absence of features in the time domain signal (see Chapter 1).

A 7-second window compares favorably to the 10-second window used in manual interpretation.

Table 13. Network depth analysis for the classification of abnormal EEGs

Configuration	# Convolutional Layers	Error (%)
$L_1 + F$	1	53.4%
$L_1 + L_2 + F$	2	22.9%
$L_1 + L_2 + L_3 + F$	**3**	**21.2%**
$L_1 + L_2 + L_3 + L_4 + F$	4	25.8%

The CNN systems that were presented thus far used 22 channels from a standard TCP

montage. The localization studies, which analyzed the system's performance at different regions

of the scalp (see Figure 16), were conducted with 4 channels for each scalp region. The results are

reported in Table 14, and show that the best performance is obtained from the occipital region

(Region IV in Figure 16). These results show that the performance for the best localized system is

worse than the performance when all the channels across the scalp are included (26.2% vs. 21.2%).

This could be attributed to a lack of training data for this very complex system.

It is interesting to compare the trends that the performance exhibited for both, the HMM

Table 14. Abnormal EEG classification based on scalp location of the input channels

Region	Channels	Error (%)
REGION I	Fp1-F7, Fp1-F3, Fp2-F4, Fp2-F8	42.7%
REGION II	T3-C3, C3-Cz, Cz-C4, C4-T4	30.1%
REGION III	T3-T5, C3-P3, C4-P4, T4-T6	53.4%
REGION IV	T5-O1, P3-O1, P4-O2, T6-O2	26.2%
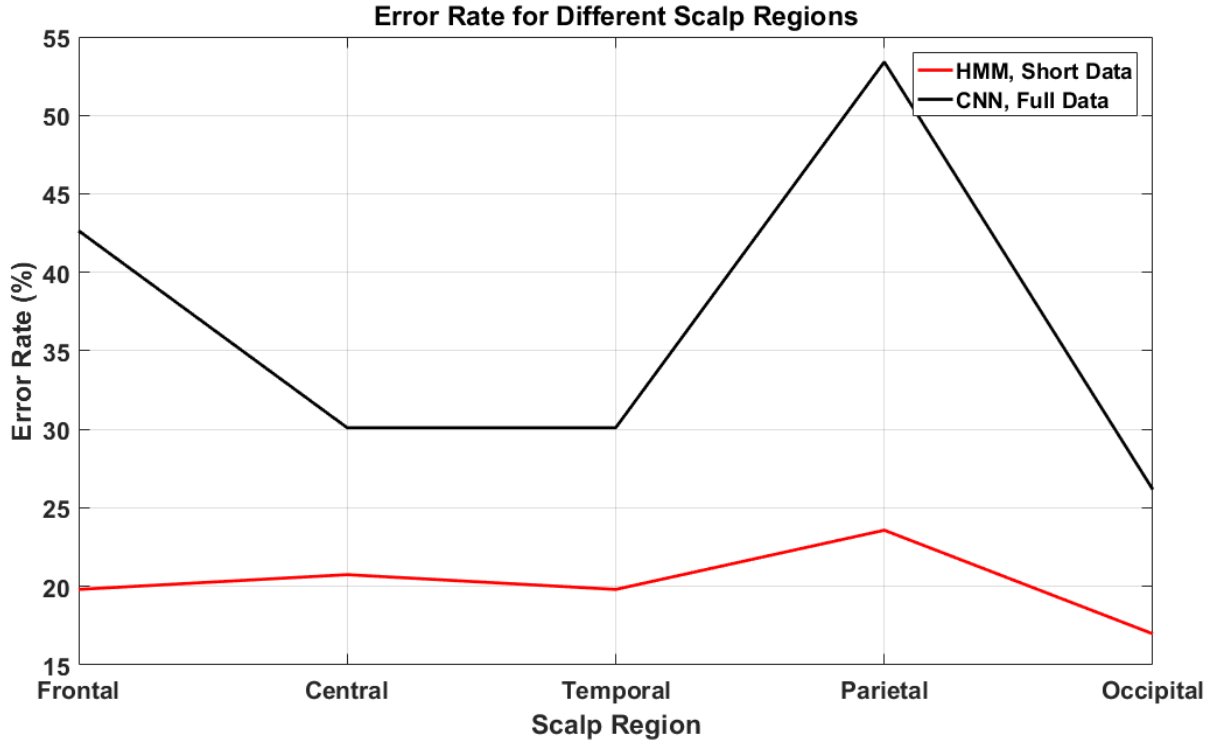
Figure 18. Performances for the HMM system with the short dataset and the CNN system with the full dataset with respect to the location of the input on the scalp.

system with the short dataset and the deep learning CNN system with the full dataset. Figure 18 shows that the behavior for both systems with regards to location of the input follows a very similar pattern, with the best results being produced by the occipital channels and the worse results being produced by inputs from the central-parietal region.

The results presented demonstrate that the best performance for the system with the full database was obtained with the CNN/MLP algorithm, which achieved a 21.2% error rate. The confusion matrix for this system is presented in Table 15. The table shows that the correct identification for normal and abnormal EEGs is more balanced for CNN-MLP than for the GMM-HMM-SdA system. While the normal EEG correct detection decreased, the abnormal correct

Table 15. Confusion Matrix for the CNN-MLP system.

Ref/Hyp	Normal	Abnormal
Normal	81.9%	18.1%
Abnormal	24.6%	75.4%

41

detection increased, contributing to a better overall performance.

4.4   Summary of Results

All of the systems that were evaluated and compared in this chapter are summarized in this section for easier comparison. In summary, some pilot experiments based on Random Forest and kNN were conducted to assess the feasibility of the problem. Then, the sequential nature of the data was exploited through the implementation of a GMM-HMM system for a short dataset. The best GMM-HMM system was then trained and evaluated on a more extensive database. This also allowed more sophisticated technology based on deep learning to be evaluated. Table 16 presents a summary of the performance obtained by these systems.

Table 16. Summary of results for the implemented abnormal EEG classification systems.

System Description	Short Dataset Error	Full Dataset Error
kNN (k=20)	41.8%	N/A
RF (Nt=50)	31.7%	N/A
PCA-HMM #GM = 3 #HMM States = 3)	25.6%	32.6%
GMM-HMM (#GM = 3 #HMM States = 3)	17.0%	26.1%
Epoch-Based HMM-Majority Vote	26.6%	24.6%
Epoch-Based HMM-SdA	27.2%	22.1%
CNN-MLP	**N/A**	**21.2%**

It is interesting to note that the performance of the systems that relied on more complex models, such as HMM-SdA, showed better performance when trained and evaluated with the full dataset. This behavior underscores the fact that construction and proper training of more complex models requires big data resources in order to achieve significant improvements in performance.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

Classifying abnormal EEGs by only considering the background information is not a trivial task. The background of a normal EEG has enough variability to make this problem particularly difficult to solve through automatic approaches. The age, state (asleep, drowsy, awake, etc.), medications and specific condition of the patients are all factors that affect different characteristics of the background signals (Ebersole & Pedley, 2014). This study showed, however, that modern deep learning techniques can use the background EEG to make decisions about the normality of the record with error rates as low as 21.2%.

All of the systems that have been compared in this study have been trained and evaluated with data labeled as either normal or abnormal. In this sense, the systems are expected to automatically model all the conditions that define an EEG as abnormal, and learn the best data representations (from extracted features) for this particular problem. Deep learning techniques have proven to be successful at discerning the best representations for problems such as speech and image recognition (Szegedy et al., 2015), (Sainath et al., 2013). However, these networks require a large number of parameters, deep networks and big data to provide state of the art performance.

The results presented in Chapter 4 show that the performance of the CNN-MLP system dramatically improves as the network becomes deeper. It is also possible to observe, however, that the performance starts decreasing after 3 layers. From this behavior, it can be inferred that the training data set used in this study is still not large enough to train networks deeper than 3 layers in conventional ways. Our models operate with a considerably lower number of parameters than state-of-the-art systems. Conversely, it can be argued that the training error obtained by the CNN-

MLP system is not reaching an optimal level, with the current dataset, and therefore, rather than more data, better algorithms and different training techniques should be explored. While this is a valid research path, it is also important to consider the information revealed by the confusion matrices presented in the previous chapter.

The best CNN-MLP system currently classifies 24.6% of the abnormal EEGs as normal (false negatives). Even though this is a good improvement from the 34.6% HMM-GMM false negative rate, the number is still relatively high, showing poor modeling of abnormality in the signals. An analysis of the misclassified files showed that the leading cause of error was related to the EEG variants that occur during sleep. The main cause being Positive Occipital Sharp Transients of Sleep (POSTS), which are physiological sharp waves that occur during stage 1 and 2 sleep and may give the false appearance of rhythmic epileptiform activity (Ebersole & Pedley, 2014). Examples that involve these transients in the training and evaluation sets introduce confusion into the systems. From these observations, we recommend a future study to additionally detect the stages of sleep and drowsiness of the subjects prior to abnormal EEG classification.

The scalp location experiments show that the observation that occipital channels are better for normal/abnormal EEG discrimination generalizes across models. The features obtained from occipital channels seem to have more discriminative power for the abnormal classification of EEGs. In this sense, models that operate with more data from occipital channels could potentially exceed the results presented in this thesis.

The purpose of the present study is to establish a baseline for the automatic classification of abnormal EEGs. Several models have been compared and analyzed. However, there are still many aspects of the problem that are beyond the scope of this thesis. Our initial approach to modeling abnormality in an EEG was largely based on successful approaches in speech and image

recognition. For this reason, several HMM systems were explored. The deep learning part of this study focused mostly on CNN implementations. However, in speech recognition, recurrent architectures have shown great promise in acoustic modeling (Graves et al., 2013). For example, Long Short-Term Memory Recurrent Neural Networks (LSTM RNNs) (Sak et al., 2014) would possibly present an improvement over the system presented here.

In this thesis, a baseline system that automatically identifies abnormal EEGs has been presented. Several models, from non-parametric models such as kNN and RF to parametric HMM-GMM models were implemented and compared. An end-to-end deep learning approach based on CNN-MLP was then implemented with a larger dataset and compared to the HMM-GMM models. The best error rate on the large dataset was 21.2%. This study demonstrated the feasibility of the automatic abnormal EEG classification only through the consideration of short amounts of EEG background signal. The data associated with this study is available in the public domain with the hope that future research will extend these baseline results.

REFERENCES

Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533–1545.

ACNS. (2006). *Guideline 6: A Proposal for Standard Montages to Be Used in Clinical EEG [White Paper]. Retrieved from*. Retrieved from http://www.acns.org/pdf/guidelines/Guideline-6.pdf

Azuma, H., Hori, S., Nakanishi, M., Fujimoto, S., Ichikawa, N., & Furukawa, T. A. (2003). An intervention to improve the interrater reliability of clinical EEG interpretations. *Psychiatry and Clinical Neurosciences*, *57*(5), 485–489.

Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., & Greenspan, H. (2015). Chest pathology detection using deep learning with non-medical training. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. New York, New York, USA: IEEE.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, *41*(1), 164–171.

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis & Machine Intelligence*.

Bishop, C. (2011). *Pattern Recognition and Machine Learning* (2nd ed.). New York, New York, USA: Springer.

Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. *Proceedings of COMPSTAT'2010*, 177–186.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Cao, C., Tutwiler, R. L., & Slobounov, S. (2008). Automatic classification of athletes with residual functional deficits following concussion by means of EEG signal using support vector machine. *IEEE Transactions on Neural Systems & Rehabilitation Engineering*, *16*(4), 327–335.

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 349–356.

Chu, C., Belavý, D. L., Armbrecht, G., Bansmann, M., Felsenberg, D., & Zheng, G. (2015). Fully Automatic Localization and Segmentation of 3D Vertebral Bodies from CT/MR Images via a Learning-Based Method. *Public Library of Science One*, *10*(11).

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv*, 1–9. Retrieved from http://arxiv.org/abs/1412.3555

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*, 3.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification*. *New York: John Wiley, Section*.

Ebersole, J. S., & Pedley, T. A. (2014). *Current practice of clinical electroencephalography* (4th ed.). Philadelphia, Pennsylvania, USA: Wolters Kluwer.

Fernandez-Varela, I., Hernandez-Pereira, E., Alvarez-Estevez, D., & Moret-Bonillo, V. (2017). Combining machine learning models for the automatic detection of EEG arousals in polysomnographic recordings. *Computers in Biology and Medicine*, *87*, 77–86.

Finnigan, S., & van Putten, M. (2013). EEG in ischaemic stroke: quantitative EEG can uniquely inform (sub-)acute prognoses and clinical management. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *124*(1), 10–19.

Gales, M., & Young, S. (2007). The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends® in Signal Processing*, *1*(3), 195–304.

Girshick, R. (2016). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* (Vol. 11–18–December, pp. 1440–1448).

Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep Learning* (1st ed.). Cambridge, MA, USA: MIT Press.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *Acoustics*, 6645–6649.

Harati, A., Golmohammadi, M., Lopez, S., Obeid, I., & Picone, J. (2015). Improved EEG event classification using differential energy. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (pp. 1–4). Philadelphia: IEEE.

Harati, A., Lopez, S., Obeid, I., Jacobson, M., Tobochnik, S., & Picone, J. (2014). The TUH EEG Corpus: A Big Data Resource for Automated EEG Interpretation. In *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–5). Philadelphia, Pennsylvania, USA.

Hau, D., & Chen, K. (2011). Exploring Hierarchical Speech Representations Using a Deep Convolutional Neural Network. In *11th UK Workshop on Computational Intelligence (UKCI)* (pp. 1–6). Manchester.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., … Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River, New Jersey, USA: Prentice Hall.

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

Keysers, D., Deselaers, T., Gollan, C., & Ney, H. (2007). Deformation Models for Image Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(8), 1422–1435.

Kingma, D. P., & Ba, J. L. (2015). Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations 2015*, 1–15.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Lee, H., Pham, P., Largman, Y., & Ng, A. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems* (pp. 1–9).

Lodder, S. S., & van Putten, M. (2013). Quantification of the adult EEG background pattern. *Clinical Neurophysiology*, *124*(2), 228–237.

Lopez, S., Gross, A., Yang, S., Golmohammadi, M., Obeid, I., & Picone, J. (2016). An Analysis of Two Common Reference Points for EEGs. In *IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–4). Philadelphia, Pennsylvania, USA.

Lopez, S., Suarez, G., Jungries, D., Obeid, I., & Picone, J. (2015). Automated Identification of Abnormal EEGs. In *IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–4). Philadelphia, Pennsylvania, USA.

Mohamed, A. (2014). *Deep Neural Network Acoustic Model for ASR*. University of Toronto. Retrieved from https://tspace.library.utoronto.ca/bitstream/1807/44123/1/Mohamed_Abdel-rahman_201406_PhD_thesis.pdf

Mohamed, A., Dahl, G. E., & Hinton, G. (2012). Acoustic Modeling Using Deep Belief Networks. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 14–22.

Muthukumaraswamy, S. D. (2013). High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Frontiers in Human Neuroscience*, *7*.

Obeid, I., & Picone, J. (2016). The Temple University Hospital EEG Data Corpus. *Frontiers in Neuroscience, Section Neural Technology*, *10*, 196.

Obeid, I., & Picone, J. (2017). Machine Learning Approaches to Automatic Interpretation of EEGs. E. Sejdik & T. Falk (Eds.), *Biomedical Signal Processing in Big Data* (1st ed., p. N/A). Boca Raton, Florida, USA: CRC Press.

Pascanu, R., Gülçehre, Ç., Cho, K., Bengio, Y., Gulcehre, C., Cho, K., & Bengio, Y. (2013). How to Construct Deep Recurrent Neural Networks. *CoRR*, *abs/1312.6*, 1–10. Retrieved from http://arxiv.org/abs/1312.6026

Peker, M. (2016). An efficient sleep scoring system based on EEG signal using complex-valued machine learning algorithms. *Neurocomputing*, *207*, 165–177.

Picone, J. (1990). Continuous Speech Recognition Using Hidden Markov Models. *IEEE ASSP Magazine*, *7*(3), 26–41.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(January), Appendix 3A.

Robinson, T., & Fallside, F. (1991). A recurrent error propagation network speech recognition system. *Computer Speech and Language*, *5*(3), 259–274.

Sainath, T. N., Mohamed, A., Kingsbury, B., & Ramabhadran, B. (2013). Deep convolutional neural networks for LVCSR. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8614–8618).

Sak, H., Senior, A., & Beaufays, F. (2014). Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. *Interspeech 2014*, (September)

Saon, G., Sercu, T., Rennie, S., & Kuo, H.-K. J. (2016). The IBM 2016 English Conversational Telephone Speech Recognition System. In *Proceedings of the Annual Conference of the International Speech Communication Association* (Vol. 08–12–Sept, pp. 7–11).

Scheuer, M. L., Bagic, A., & Wilson, S. B. (2017). Spike detection: Inter-reader agreement and a statistical Turing test on a large data set. *Clinical Neurophysiology*, 128(1), 243–250.

Smith, S. (2005). EEG in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery, and Psychiatry*, *76* (Suppl 2), ii2-ii7.

Steve Young. (1996). A review of large-vocabulary continuous-speech. *IEEE Signal Processing Magazine*, *13*(5), 45.

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 7-12–6, pp. 1–9). IEEE.

Vincent, P., & Larochelle, H. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion Pierre-Antoine Manzagol. *Journal of Machine Learning Research*, *11*, 3371–3408.

Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, *10*, 207–244.