SNAPPY TITLE GOES HERE

A Dissertation
Submitted to
the Temple University Graduate Board

---

in Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
Christian Radcliffe Ward
May / August / December, 20XX

Examining Committee Members:

Dr. Your Advisor , Advisor, Dept. of Electrial and Computer Engineering
Dr. Member One, Dept. of Z and X
Dr. Member Two, Dept. of Z and X
Dr. Member Three, Dept. of Z and X
Dr. Member Four, External Reader, Dept. of Z and X

ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque

tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

ACKNOWLEDGEMENTS

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Words.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AAC American Academy of Clinicians

ABPN American Board of Psychiatry and Neurology, Inc.

ACNS American Clinical Neurophysiology Society

AD Alzheimer's Disease

ADHD attention-deficit/hyperactivity disorder

ADJUST Automatic EEG artifact Detection based on the Joint Use of Spatial and Temporal features

ANN Artifical Neural Network

ApEn Approximate Entropy

BW Baum-Welch

BCI brain-computer interface

BSS blind source separation

CHB Children's Hopsital of Boston Massachusettes Institute of Technology Scalp EEG Database

CD Cosine Distance

CEP Cepstral Coefficient

COH spectral coherence

CRR Correct Recognition Rate

CRIM Centre de Recherche d'Informatique de Montreal

CSP common spatial pattern

DBN Deep Belief Network

DET Detection Error Tradeoff

DNN deep neural network

DP Dircihlet Processes

DT Decision Tree

EC Eyes Closed

ECG electrocardiogram

ED Euclidean Distance

EDF European Data Format

EEG electroencephalogram

EER equal error rate

EM expectation maximiation

EMD emperical mode decomposition

EMG electromyography

EO Eyes Opened

EOG electrooculography

ERP evoked response potential

ET epileptiform transient

FA factor analysis

FAR false acceptance rate

FASTER Fully Automated Statistical Thresholding for EEG artifact Rejection

FRR false rejection rate

FSC Fuzzy Sugeno Classifier

FFT Fast Fourier Transform

GFP global field potential

GMM Gaussian Mixture Model

GPED generalized periodic epileptiform discharge

GMM-UBM Gaussian Mixture Model-Universal Background Model

GMMHMM Gaussian Mixture Model based Hidden Markov Model

HDP Heirarchical Dirichlet Process

HMM Hidden Markov Model

HTER half total error rate

HTK Hidden Markov Toolkit

ICA independent component analysis

ICU Intensive Care Unit

iEEG intracranial electroencephalogram

IMF intrinsic mode function

I-Vector Identity Vector

JFA joint factor analysis

KNN K-Nearest Neighbor

LDA Linear Discriminate Analysis

LS-SVM Least Squares Support Vector Machine

LSTMNN Long Short-Term Memory Neural Network

LMBPNN Levenberg-Marquardt Backpropagation Neural Network

LOOCV leave one out cross validation

MAP maximum a priori

MCI mild cognitive impairment

MD Mahalanobis Distance

ML Machine Learning

MLE maximum likelihood estimation

MLPNN multilayer perceptron neural network

MFCC Mel Frequency Cepstral Coefficient

MSR Microsoft Research

mMSE modified multiscale sample entropy

NBC Naive Bayes Classifier

NEDC Neural Engineering Data Consortium

NN Neural Network

PCA principal component analysis

PD periodic discharge

PLDA probabilistic linear discriminant analysis

PLED periodic lateralized epileptiform discharge

PLI phase lag index

PMean pooled mean

PhysioNet Database PhysioNet EEG Motor Movement/Imagery Database

PNN Probabilistic Neural Network

PSD Power Spectral Density

QDA Quadratic Discriminant Analysis

RA1 Research Aim 1

RA2 Research Aim 2

RA3 Research Aim 3

RBFNN Radial Basis Functional Neural Network

REM random eye movement

RF Random Forest

RMS Root Mean Squared

SampEn Sample Entropy

SOM self organizing map

SPMD Single Program Multiple Data

SVM Support Vector Machine

TBR theta beta ratio

TCP Trans-Cranial Parasagittal

TUH Temple University Hopsital

TVM total variability matrix

TUH Corpus Temple University EEG Corpus

UBM Universial Background Model

VEP Visually Evoked Potential

WPD wavelet packet decomposition

Chapter 1

INTRODUCTION

The ability to communicate underlies the major functions of the brain. Given the array of tools at our disposal (voice, facial expressions, hands, feet and eyes) our ability to communicate is limited only by our inventiveness. However, this system of communication limits our brain by forcing it to indirectly communicate through these tools. When we wish to study the brain itself problems arise because the majority of measurements come through indirect means. This is further complicated as the ideas to be expressed become more complex either in terms of emotional context or severity, such as pain and illness.

Presently, electroencephalography is the principle method of directly communicating with the brain. While the communication is one directional, in that we can only listen, it affords opportunities not available through our human faculties. electroencephalograms (EEGs) may be used to discern the incidence of epilepsy and stroke [1], study neural responses to stimuli [2], or even neural control feedback [3]. Recently, the advent of inexpensive commodity-grade EEG headsets [4] has expanded the field to include areas such as gaming, neuro-modulation, and mindfulness training [5].

These advances allow for direct and more timely interpretation of EEGs via the creation of digital signal processing tools that can identify or predict neural activity [6]. In clinical settings this technology assists neurologists in reviewing long recordings [7], communicating with patients [2], and processing artifacts [8]. These tools leverage multidimensional statistical models [9, 10, 11] to enhance our understanding of

EEGs. In research settings, this technology has facilitated advances in brain-computer interfaces (BCIs) [12] and seizure prediction [13].

Historically, computer-based EEG interpretation has been only moderately effective despite large quantities of research [14, 15]. One key problem is that brain function (and by extension an EEG recording) is highly variable, requiring very large sample sizes in order to create robust statistical models [16]. The most powerful statistical methods generally require even larger samples sizes to assure convergence [17]. Until recently it has been difficult to collect, store, and process such large EEG datasets.

Modern digital data collection methods, in both clinical and research settings, have made 'big neural data' feasible [18]. However, these datasets must be *annotated* prior to being useful for training statistical models. Annotated data is produced when an expert reviews the recordings by marking which segments of the recordings correspond to known phenomena [19]. These annotations can be at the macro scale (such as 'seizure') or the micro scale (such as 'sharp spike wave'). Not surprisingly, EEG annotation is manually intensive making it rarely cost effective to ask clinicians to perform it at a fine-grained level [20]

There are communication problems between even well trained clinicians on how and what to annotate on recordings. This is evident by moderate consensus agreement when annotating simple events such as variations of spike waveforms [21, 22, 23, 24]. Conflicting annotations make it difficult to produce 'gold standards' of annotations used for training new clinicians and for leveraging the power of *supervised* Machine Learning (ML) techniques.

Supervised ML techniques rely on this annotated data, more commonly called *labeled* data within the ML community, to produce sufficient models of known classifications. By using prior knowledge of the data, models can be trained to classify

previously unseen data in classes such as background, seizure, and sleep. However, a common problem with these techniques is a lack of strong consensus for each class [6]. Thus the system is inherently limited by the quality and quantity of its prior knowledge.

The difficulty increases when building *unsupervised* ML techniques that operate on unlabeled data [20, 25]. Now the techniques are tasked with first determining how to partition the data into classes and then performing classification on those self-generated labels. This typically requires additional data beyond a supervised approach, but removes the stipulation of prior knowledge.

Despite the majority of research focusing on supervised ML, an unsupervised ML method may best suited for interpretation of EEGs. Unsupervised approaches are decoupled from clinicians because there is no need for labeled data. Clinicians are capable annotators, but even in their area of expertise they have biases which manifest in poor inter-rater agreement when aggregating annotations [26]. Furthermore, as the use cases of EEGs grow they advance beyond what clinicians typically annotate, meaning it is impossible to provide a documented ground truth. Given such constraints, this work introduces Identity Vectors (I-Vectors) as an unsupervised machine learning method for EEGs with the aim of supplanting the reliance on clinician annotations.

1.1 The Landscape of Electroencephalograms

Before outlining the aims of this work, a brief background is provided to a shared understanding of the relationships between EEGs, algorithms and clinicians. Chiefly among these relationships is the way in which algorithms and clinicians are trained and perform annotations. Specific attention is paid to how clinicians, as individuals

and groups, produce the annotations used for algorithm development. The performance of these algorithms is outlined to contrast with the scope and performance of their human counterparts. Attention is focused on the algorithms areas of application and performance.

1.1.1 Clinician Development

Clinicians undergo extensive training, often culminating in a fellowship to specialize in the treatment of epilepsy, sleep disorders, or intensive care. These specializations require the ability to interpret EEG recordings[1] for which the clinician can be certified through the American Board of Psychiatry and Neurology, Inc. (ABPN). The American Academy of Clinicians (AAC) works with the ABPN to ensure clinicians are adequately trained, but cautions that "[N]ot all hospital credentialing boards require sub-specialty training to allow individuals to interpret EEGs"[2]. Sub-specialty certifications are limited to topics such as brain injury, neuromuscular issues, and epilepsy.

Beyond this, clinicians refine their skill on the patients they encounter through their practice of medicine. Principle among these skills is their ability to accurately annotate EEGs recordings. Annotations focus on documenting the activity of the brain via signals recorded from strategically placed electrodes extracranially (on the scalp) or intracranially (on the surface of the brain) [27]. The methodology of annotating and interpreting EEG recordings is part of the certification process, but the Epilepsy Foundation contends that "EEG training for clinicians is inadequate"[3]. In

[1]Taken from: `https://medicine.yale.edu/neurology/education/fellowships/epilepsy_eeg/`

[2]Taken from: `https://www.aan.com/uploadedFiles/Website_Library_Assets/Documents/4.CME_and_Training/2.Training/3.Fellowship_Resources/3.How_to_Apply_for_a_Fellowship/Epilepsy\%20Fellowship\%20FAQ.pdf`

[3]Taken from: `http://www.epilepsy.com/article/2014/12/eeg-training-clinicians-inadequate`

spite of this, clinical annotations remain the best tool for assessing the behavior and state of a brain [28].

Even with all their training and successful treatment of the myriad of brain disorders, clinicians are not without their inconsistencies as they are human [21]. Firstly, their ability to annotate accurately is often surpassed by the amount of data produced from tests. This leads to annotation consuming a disproportionate amount of their work hours. Secondly, their formal education ensures they are in agreement on terminology and its manifestation [22]. However, performance in consensus-bases studies suggests there are disagreements over which waveforms are of interest to each clinician [21, 23, 24].

Thus it is clear that clinicians are capable interpreters because they readily determine the correct diagnosis from a EEG recording. However, their reasoning for these assessments have the potential to be disparate. This behavior is not unique to a specific subset of conditions as it is readily apparent in the lack of annotation consensus in sleep [24], seizure [21] and cardiac [29] EEG recordings.

Even when presented with data common to their expertise, pairwise clinician similarity (Cohen's κ statistic[4]) is moderate (0.41-0.60) at best [21] and group performance varies from slight (0.0-0.20) to almost perfect (0.81-1.00) [29]. This suggests clinicians identify different, but valid, indicators of disorders. Ultimately this produces multiple divergent, but correct, sets of annotation from one dataset. While not problematic for diagnosing disorders, it makes it difficult to develop ML algorithms when there are multiple 'right' answers.

[4]The statistic is not perfect [30], but does appear to be among the most common reported in studies assessing neurologist performance.

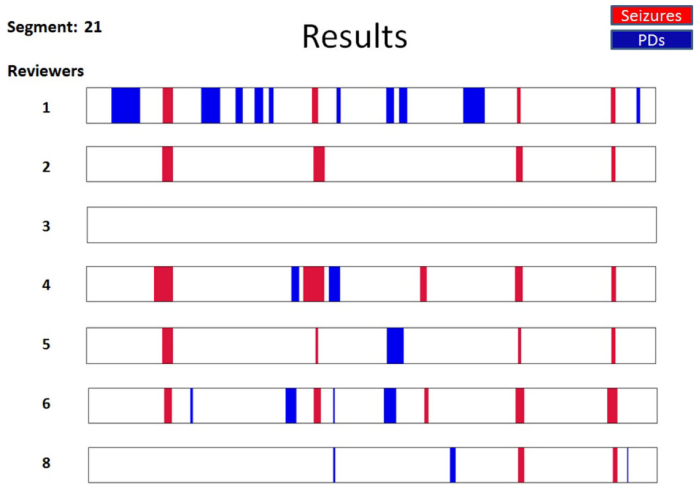


Figure 1.1: In Halford et al. [23], seven reviewers were asked to annotate for seizures and PDs. The annotation results of the hour long recording, Segment 21, show that six reviewers labeled seizure events, five labeled PDs, and one labeled nothing. The quantity of annotation varies as does the spatial alignment between between reviewers.

1.1.2 Clinical Annotations

The ability to produce correct annotations is a fundamental component of EEG based research. In order to validate the performance of algorithms, clinicians must provided annotated data. These datasets are annotated through the lenses of the clinician's specialization and the patient's condition or diagnosis. As discussed previously, even when annotating the same data, clinicians struggle to come to consensus about its contents. Figure 1.1 shows the results of seven clinicians annotating an hour long segment for seizures and periodic discharges (PDs). Nearly all the clinicians annotate abnormal events, save one, but the diversity and quantity of annotations are inconsistent.

Further complicating matters is that investigators often produce their own datasets, specifically for a given study. This occurs because existing datasets lack annotations, subject information, recording parameters, or protocols necessary to address their

specific research questions. This makes it difficult to reuse previously annotated data because there is nothing is standardized. While one study annotates the other two do not, and then all three present with different sampling rates, recording durations, and electrode layouts.

While these decisions are practical with respect to specific studies, this behavior prevents supervised ML techniques for being applied across datasets. Without consistent sampling rates, the datasets may need to be interpolated to produce consistent windows of data. Mismatches in electrodes, inconsistent annotations, and artifacts are often manually resolved via the experiment team's limited knowledge or by possibly requiring the assistance of yet another clinician. While algorithms may overcome noise inherent in the data, this is only possible if there is a plethora of well annotated data from which to learn.

Annotations start, as shown in Figure 1.2, as waveforms whose variations conform to similar behaviors. Not all annotations are related to medical conditions, as eye blinks and background are often considered to be noise. Differentiating such noise from waveforms of interest, like generalized periodic epileptiform discharges (GPEDs), periodic lateralized epileptiform discharges (PLEDs), spike and sharp wave complexes, and triphasic waves, is a critical step in reading an EEG. The American Clinical Neurophysiology Society (ACNS) defines an exhaustive list of EEG terms, including background characteristics, which are outlined in [29]. Clinicians are well versed in the terminology, but struggle in their ability to accurately match waveforms with appropriate labels [31].

The waveform examples from Wulsin et al. [14] are drawn from a seizure dataset. However, the waveforms are not unique to seizure recordings and could also be found in any of the other active EEG research fields such as attention/workload measurement [32], biometric identification [33], BCIs [5], evoked response potentials (ERPs)

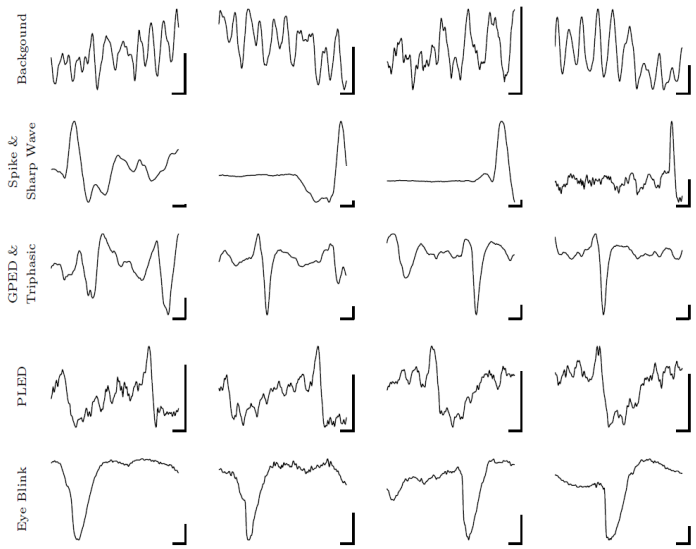


Figure 1.2: Annotations used for the work of Wulsin et al in [14]. Notice the placement of the spike does not need to precede or succeed the sharp wave. GPED and PLED typically occur over a range of channels making them context dependent.

[34], and sleep stage classification [35]. Each field focuses on different facets of an EEG recording and may have distinct waveforms. Other sources for distinct waveforms include subject related traits, such as their age [24, 36] and genetics [37].

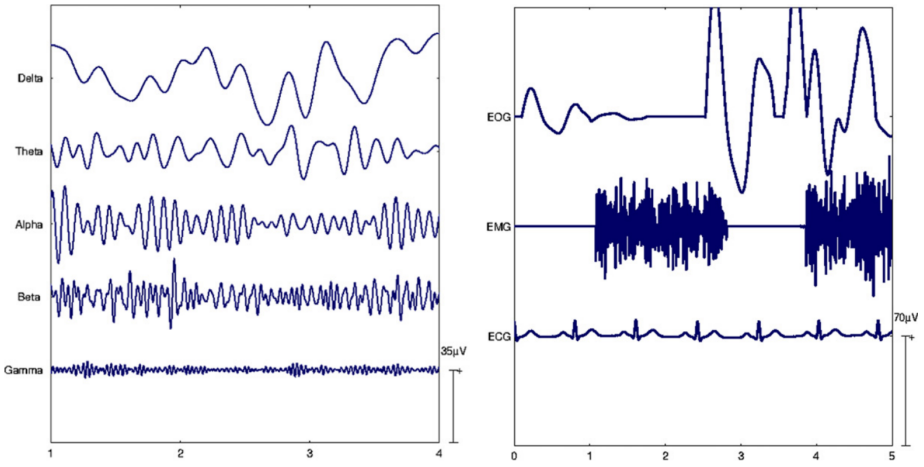
In summary, the fundamental technical challenge of training robust algorithms for automatic EEG interpretation is the diversity of annotated data. Seizure data differs from ERP data which differs from sleep data, making it difficult, if not impossible, to find clinicians capable of accurately annotating all of it. The lack of large diverse sets of thoroughly annotated data encumbers the advancement of algorithm based annotators/classifiers. This is exemplified by the struggle to develop ML algorithms capable of meeting performance levels deemed acceptable by clinicians and the inability to produce consistent universal ML classifiers.

1.1.3 Algorithm Applications

While major research avenues align with clinical applications (sleep, seizure, and various brain disorders), the use of ML provides avenues for novel applications as research progresses such as BCI, biometric verification, ERPs, and brain state workloads. Despite the variety of unique classification tasks, they all face similar fundamental performance hurdles. Chiefly among these are the necessary steps of pre-processing to address artifacts and production of acceptable feature sets. Within a given EEG recording it can be necessary to address the background waveforms that comprise the majority of the datasets.

EEG artifacts are often hard to classify because they appear as waveforms that resemble, Figure 1.2, the more critical spikes and sharp waves of seizures [26] or the natural brain frequency rhythms [38], Figure 1.3. While artifacts impact clinicians and algorithms, selection of an optimal feature set is unique to the algorithms. This is because feature sets are often paired with the type of EEG data being classified. The result is a wide range of potentially useful features consisting of but not limited to Power Spectral Density (PSD) features [39], spatial and temporal features [40], Cepstral Coefficient (CEP) feautres [41], auto-regressive parameters [42], and normalized raw data [43].

Despite focusing on waveforms of interesting via artifact correction and feature selection, the majority of EEG often consists of background signals[14, 44, 38]. This is frequently a problem for rare events like seizures, but is a boon to subject verification tasks and the biometric community [45]. Additionally, there are many less studied conditions that manifest throughout a recording, such as alcoholism [46], emotional state [47], pain [9], and mental focus/workload [48].



(a) Brain Rhythms (b) Artifacts

Figure 1.3: Example of artifacts (b) similarity to the natural rhythms of the human brain (a) used for the work of Uriguen et al in [38].

To motivate the implication of these areas of research a brief review of six common EEG classification fields is presented. The use of algorithms for seizure, sleep, BCIs, ERPs, and mental/workload classification are readily associated with clinician driven research, while EEG based biometrics branch out beyond their well defined knowledge base.

Seizures A substantial portion of work in this field focuses on correctly identifying and locating seizures [49, 50, 51, 52]. By isolating seizure events, researchers can focus on the properties of the seizure for the purposes of classification and waveform modeling [14, 53, 54]. The knowledge gained in this process makes it possible to predict seizures in real-time [6, 13]. Seizure events are typically high energy and frequency wavefroms with synchronization across channels [23].

Sleep Studies Sleep state classification labels the transition from wakefulness to random eye movement (REM) sleep. Sleeping EEG recordings are often cleaner due to lack of movement artifacts which improves their clarity for clinicians and reduces

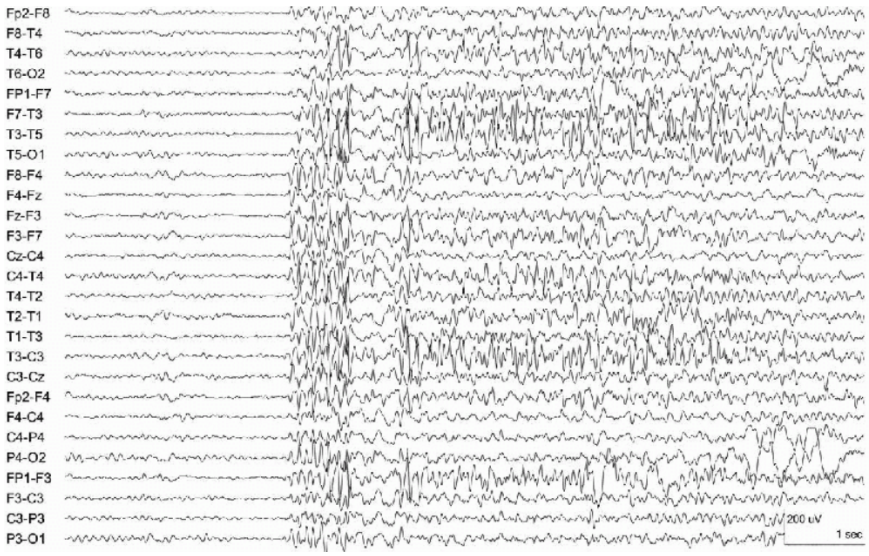


Figure 1.4: A segment of an EEG recording taken from a subject at the onset of a generalized seizure. Note that after the seizure starts, activity is not uniform across all channels. Image sourced from Tatum and Tatum[55].

pre-processing for algorithms [56]. Despite this and a closed set of distinct stages, sleep stage classification suffers from inter-rater agreement problems [24]. Sleep spindles and K-Complexes serve as the main indicators of sleep along with pronounced changes in band Power Spectral Density [57]. While seizures often manifest during sleep, other issues can also be addressed such as sleep apnea [35] and overall brain functionality/health [58].

Biometrics Multiple studies have focused on the use of EEGs to identify and verify subjects, irrespective of any associated disease and disorder [59]. The results of such work suggest that individuals have distinct EEG fingerprints [60, 61, 62] which may relate to potential inheritable characteristics [61, 63]. A major theme in biometrics is understanding how different brain states impact these fingerprints. The work of Rocca et al. showcases brain distinctiveness when using a common testing state of resting eyes closed [33], spectral coherence as discrimination feature [64], and

Figure 1.5: A segment of an EEG recording taken from a subject in the second of phase sleep. Note the present of sleep spindes, black arrow, across multiple channels. Image sourced from Tatum and Tatum[55].

techniques to reduce the feature set into sparse mappings [65]. Some approaches overlap with other applications by invoking response potentials [66], focusing on specific brains states of sleep [67], or restful states with eyes open and closed [68]. Even the longitudinal stability of biometric EEGs is tested [69] to determine viability for long term applications.

Brain Computer Interfaces BCI technology finds ways to get information into and out of a brain. The most advanced applications of this are restoring functionality to those unable to use their body [70, 71]. This requires algorithms robust to changes in subjects, but sensitive to spatial and temporal facets of EEG recordings [72, 73]. Development of subject invariant algorithms has led to disparate training protocols with transfer learning using multi-subject models [74] and zero-calibration training being subject specific [75]. This leads to a similar problem as sleep, where the wave-

forms are well understood, but their manifestation across populations complicates their performance.

Evoked Response Potentials ERPs are a stimulus response and not a voluntary action. A well documented case of ERP is the P300 signal that triggers in the parietal/occiptal region 300 milliseconds after seeing an image of interest [2]. This signal is commonly used to enable subjects to communicate via P300 spellers. These spellers flash the alphabet before a subject waiting for a letter of interest to trigger an ERP, which allows them to build words [76]. This approach allows a brain to communicate without the need of a body, but also has applications for testing processing time of visual and auditory stimulus response [77].

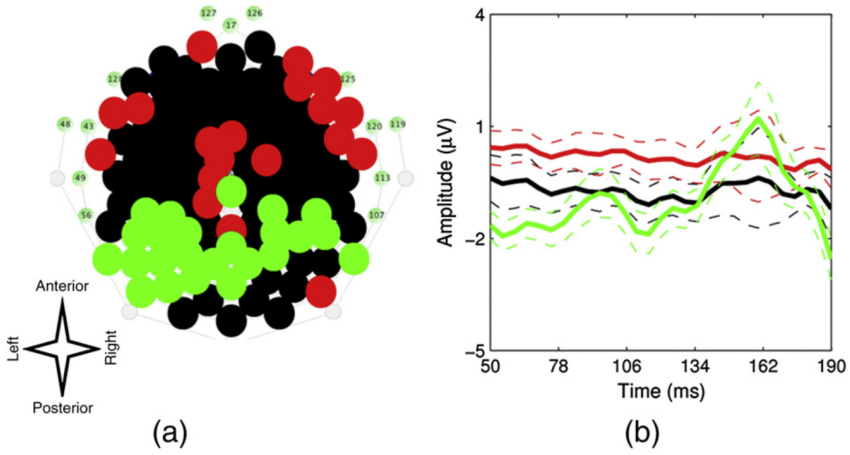


Figure 1.6: A 2D mapping of the electrodes and their group averaged waveform (solid lines). The standard deviations of the channel averaged are given as the dashed lines. Image sourced from Karamzadeh et al.[77].

Brain State/Workload Analysis of involuntary conditions address the state of a person's brain which can refer to the emotional state, disease state, or attention/workload state. Those afflicted with Alzheimer's [78], alcoholism [46], and mental disorders such as attention-deficit/hyperactivity disorder (ADHD) and Bi-Polar

disorder [79] present with distinct EEG features. Knowing these conditions can manifest in the EEG recordings provides context for the how the known underlying biological changes alter a subject's EEGs. This is exemplified by studies measuring how stress impacts cognitive function [80] and a brain's workload during attention dependent tasks [48].

1.1.4 Algorithm Development

The development of ML techniques for EEG tends to focus on areas well understood by clinicians, detecting seizures [13, 14], identifying the stages of sleep [35, 81], capturing ERPs [34, 82], or processing BCI signals. Minimal focus has been given to a generalized classifier for interpreting multiple types of EEGs [83]. The approach closest to this goal is the use of EEGs for biometrics given that subject verification works on variety datasets with similar results [84, 85, 86]. While conditional classification techniques (seizure detection, sleep classification, BCIs, and subject verification) are capable, they fail to increase our overall understanding of EEGs.

Despite the lack of a generalized classifier, the data specific classifiers rely on some amount of data pre-processing. This is necessary to address recording artifacts [11, 87, 88], optimize the available channel data [64], or generate an acceptable feature set [89]. In carrying out one or more of these pre-processing steps a preliminary amount of dimensionality reduction is introduced which becomes more pronounced as the data is windowed into epochs for a given algorithm [90, 91, 92].

Unfortunately all these steps are often unique to the type of EEG being classified which means there is no well defined protocol of feature set that applies universally. For example, seizure algorithms typically process data in windows on the order of 10s of seconds [93]. Biometric algorithms utilize channel subsets to verify a subject [45]. BCIs use spatial filters to target the regions of the motor cortex [94]. ERPs focus

on the occipital region where recognition of stimulus is triggered [72]. Things are further complicated by the varying performance within a dataset based upon subject or recording variation seen in BCI tasks [90, 73, 95], seizure recordings[6, 14, 96], and even biometric protocols [97, 98]. Due to this a comprehensive feature set remains elusive, but data specific feature sets have shown promise when paired with various algorithms.

These approaches leverage knowledge gained from the study of EEGs which makes them *domain knowledge*. Unfortunately domain knowledge comes from clinicians which means, as outlined previously, there are limits to its impact. It is critical in understanding artifacts and background (Figure 1.2), seizures (Figure 1.7), and sleep patterns (Figure 1.8) [99, 57], but clinicians have minimal knowledge specific to biometrics [84]. Thus some approaches are bootstrapped by domain knowledge, but it furthers a Catch-22. Algorithms are made dependent clinician supplied insights when the algorithms task is to provide annotations to assist those same clinicians.

Within this loop of clinician annotations driving the development progress of algorithms, is the closed set of available EEG datasets. Aside from the PhysioNet EEG Motor Movement/Imagery Database (PhysioNet Database) [100] and BCI competition [101] databases, much of the research is conducted on specific single use datasets [14, 92, 102]. Furthermore, when the PhysioNet Database database is used it is often the only dataset [86, 103, 68, 104]. There are studies that combine datasets, but that they tend to focus on biometric applications [105].

This lack of a robust data landscape manifests as variable algorithm performance based on the dataset[11, 91] or, within the context of BCI applications, as subjects being unable to use the system making them *'illiterate'* [106, 107]. This suggests there are intrinsic problems within processing EEG data. Algorithm performance can they be viewed as being dependent on the type of data (BCI, sleep, seizure, etc),

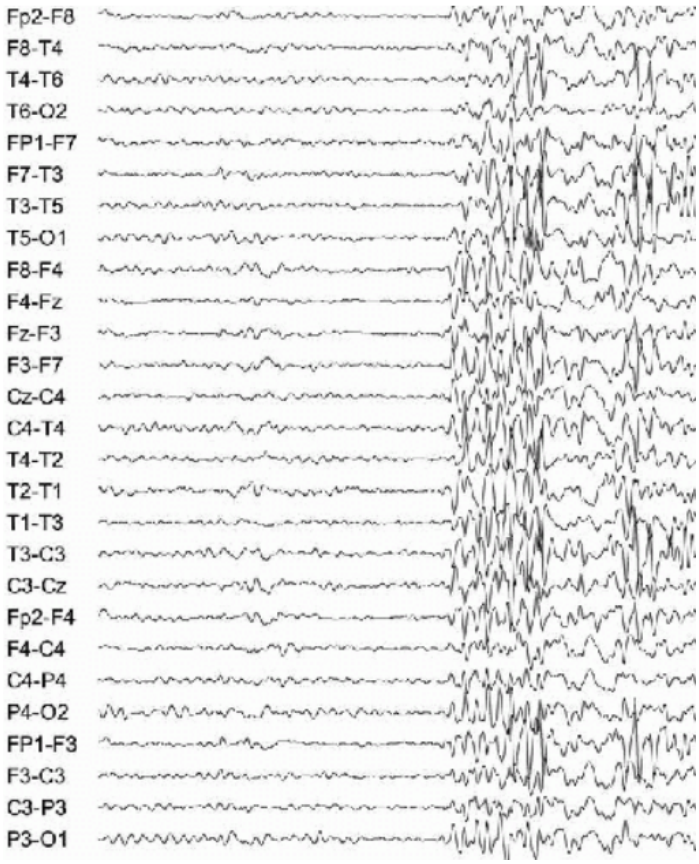


Figure 1.7: Sleep Spindle example drawn from the text of Tatum[55]. Notice the abrupt change in all recorded channels.

Figure 1.8: Sleep Spindle example drawn from the text of Tatum[55]. The black arrow indicates the location of a sleep spindle which is captured across multiple, but not all, channels.

but also the individual subjects themselves. This makes it difficult to tout any way forward given the lack of common performance benchmarks.

To address these data issues, algorithms must either be generalized, or built to tackle specific problems using domain knowledge. These two paths pair well with unsupervised (generalized) and supervised (specific) ML algorithms [34, 108, 109, 110]. The use of domain knowledge with supervised deep learning approaches have shown promise in BCI [111], sleep [112] and seizure [14, 13] classification. While their unsupervised counterparts do not require clinician support they do lag in comparative performance when addressing similar classification tasks [83, 113, 114, 110].

Efforts to address this data dependency in the BCI field produced four styles of classification schemes: *adaptive classifiers*, *matrices and tensors*, *transfer learning*, and *deep learning*[91]. Ten years on Lotte et al. [110] noted the same schemes work, but a focus should be placed on the end user. These reviews covered supervised and unsupervised algorithms indicating that no resolute classifier has been found for the BCI community. Seizure classification has progressed, but the results are often using small esoteric datasets with 5 subjects [43] and 17 subjects [13]. Biometric techniques continue to perform well, but have also failed to expand their datasets [115, 59, 86].

The potential within the EEG ML community is vast. Presently performance varies based upon dataset, feature set, algorithm, and often subject quality. While some applications have become robust, seizure, ERP, and BCI classification show promise, this field is still maturing. Given the nascent status of EEG processing borrowing techniques from an established domain may be helpful. In general much of the technology currently deployed for ML pursuits comes from the realm of speech recognition, such as the time tested Hidden Markov Model (HMM) [116]. Following this trend, the development of the unsupervised learning technique called I-Vectors could offer growth of performance and understanding for EEG classification tasks

[117]. I-Vectors are able to learn decision surfaces for the accent, age, content, gender, and language of a speaker [118]. Through a series of data modeling utilizing Gaussian Mixture Models (GMMs) [119] that produce a Universial Background Model (UBM) [120] capturing the variability of the training data in a total variability matrix (TVM), it is possible to reduce the dimensionality of various sized segments of data into robust discrimination vectors, I-Vectors [121].

1.2 Research Proposal

A clinician's primary focus is to treat their patients. Asking clinicians to produce perfectly annotated recordings to support algorithm research is not in the best interest of their patients or their overall productivity. There is little sense in asking clinicians for help to build datasets for algorithms who's goal is to reduce the time clinicians spend reading EEGs. This is clearly a Catch-22: The people that algorithms can help must first help to train the algorithms. However, clinicians do not have the time or group consensus to meet the needs of the algorithms.

The most direct solution is to find a way to annotate recordings without involving clinicians. As discussed ML-based solutions exist, but the field is diverse and lacks an apex technique. Despite the success of these techniques, fundamental problems continue to exist which must be overcome by all algorithms. These include variations in the quality of the recordings, the presence of adequate (in quality and quantity) annotated data, an acceptable feature set, and consistent channel layout across recordings. At its core the issue is identifying what characteristics of the EEG are relevant for a given classification task. In most instances, annotated data and prior knowledge is leveraged in order to reduce the dimensionality, and thus the uncertainty, in the

algorithm's classification. This approach reinforces a reliance on annotators, which is not ideal given the disparate quality and consensus of annotations.

Annotation-based techniques are presently the dominant ML approach to classifying data. This means that clinicians effectively control the algorithms' performance which makes them an external source of error. To alleviate this constraint, unsupervised ML algorithms can be developed to match the capabilities of their supervised counterparts. The benefits of equivalent performance would be significant, as unsupervised ML enables training on large diverse datasets without the need of clinicians. Countless hours of data in need of annotation could thus be labeled, producing a steady supply of data for training supervised ML algorithms and clinicians. By using I-Vectors for this process it may also be possible to uncover novel phenomena in the data similar to their use on speech signals.

1.2.1 The Research Aims

The goal of this work was to lay the foundation for an unsupervised ML system that classified and clusterd EEG recordings. The preliminary pre-dissertation work indicated it was possible for I-Vectors to perform subject verification and to sort data by similarity[5]. While promising, these results had to be expanded to determine whether I-Vectors could overcome the annotation advantage. This primarily relied on the constrained modeling processing carried out in the generation of I-Vectors. Once mastered, the process was largely transparent in its approach making it possible to study the decision surfaces used for the proscribed classification and clustering tasks.

In addition to understanding how the proposed system operated on EEGs, it was necessary to prove that I-Vectors could offer comparable performance to existing standard methods, including both ML algorithms and clinicians. However, given the

[5]See chapter 4's preliminary experiment results.

advancement of ML algorithms, the ability to cluster and verify subjects is related only to algorithms. Clinicians do perform similar tasks, but they use resources beyond EEG recordings to make their assessments such as medical reports. Thus the performance of I-Vectors was evaluated against other well documented ML of varying complexity to highlight the tradeoffs between performance, dimensionality reduction, and algorithm complexity.

From these areas of interest, three research questions were posed:

Research Aim 1: Can an I-Vector-based classification perform as well as, or better than, other applicable ML techniques?

Research Aim 2: Under what conditions does an I-Vector based system perform best?

Research Aim 3: What characteristics of EEG data do I-Vectors take advantage of in their discrimination? Is this process inherently well suited for addressing EEG classification?

By answering these questions, insight into the nature of I-Vectors and EEGs was gained. This was possible because similar I-Vectors work in the speech recognition community produced strong results related to subject verification [122], language classification [123], accent detection [118], and speaker age estimation [124]. The underlying hypothesis was that EEGs had a bounded mathematical space similar to speech signals. This space can be exploited by the constraints of the TVM which shapes the I-Vectors producing nuanced classification similar to those seen in speech.

1.2.2 The Research Experiments

The Aims of this work was addressed in three experiments: *Parameter Sweeps*, *Algorithm Benchmarks*, and *UBM-TVM Relationship*. Upon completing the experiments, the process of producing I-Vectors from EEG data was understood along with which properties of EEG and I-Vector made this approach viable for producing annotations in an unsupervised manner.

Parameter Sweeps The purpose of the *Parameter Sweeps* was to determine optimal operating parameters for applying I-Vectors to EEGs. This addressed Research Aim 2 by measuring the significance of specific features, channels, UBM mixture sizes, and the TVM training process. Testing each parameter over a range of values produced trends for a best practice approach to baseline I-Vector systems. The statistical decomposition of each dataset (abnormal, normal, motion trials, and seizure) and I-Vector development process provided background and baseline results enabling comparisons against the other published results where the data is not publicly available.

Algorithm Benchmarks In order to validate I-Vectors as an option for classification and clustering of EEG data their performance was compared against a suite of ML algorithms. The algorithms were evaluated through their sensitivity and specificity and, when applicable, their ability to cluster. These experiments addressed Research Aim 1 through a series of leave one out cross validation (LOOCV) experiments based on subject and channel classifications.

UBM-TVM Relationship The relationships between UBMs and TVMs was deconstructed to examine the trade-offs made during optimization of the TVM. Using

the reported performance of Gaussian Mixture Model-Universal Background Model (GMM-UBM) and I-Vector classifications, the influence of the mixture weighting were traced throughout the entire modeling process. This manifested as comparative feature and mixture mappings for each classification test. These mappings unlocked the fundamental statistical properties used to differentiate subjects which can then be compared across data sets as they are bounded by a common feature set. Ultimately this protocol turned I-Vectors into a powerful multi-modal signal analysis technique.

Chapter 2

BACKGROUND

Scarecrow:

The sum of the square roots of any two sides of an isosceles triangle is equal to the square root of the remaining side. Oh joy! Rapture! I got a brain! How can I ever thank you enough?

The Wizard of Oz:

You can't.

This chapter introduces the nature and use of EEGs in clinical and research settings. Clinical EEGs are used by clinicians to make diagnostic decisions in accordance with their education and training. In research settings algorithms strive to replicate the performance of clinicians through statistical modeling guided by clinician annotated data. Together these two groups are increasing our ability to discern the meaning of EEG signals.

This dissertation will examine the suitability of I-Vectors as a mathematical tool for allowing researchers to replicate clinician performance on EEGs. I-Vectors have shown promise with respect to classification and clustering of speech signals in terms of accent, age, context, gender, and language via its feature transformation process. This type of discrimination would be beneficial to understanding the phenomena that produce EEG waveforms. The I-Vector technique is introduced in depth along with the necessary criteria to evaluate it against other algorithm based discrimination techniques.

2.1 Electroencephalograms

An EEG records the electrical activity of the brain. The captured voltage signals represent the firing of neurons involved with all aspects of a brain's functionality. Through the use of EEGs we can see how the brain functions on an operational level [47], interprets stimuli [62], and changes due to diseases and disorders [46]. The applications of EEGs are primarily limited by the ability to link recorded activity to the underlying physiological condition.

A clinician's ability to annotate EEG recordings utilizes their knowledge of the relationship between waveforms and physiological conditions. An accurate diagnosis cannot be made from waveforms only as the clinician must consider the subject's history and the recording conditions of the EEG. In many cases spatial and temporal properties must be considered when assess for specific conditions related to different regions of the brain and similarities between waveforms.

Depending on application, EEG signals require radically different signal processing techniques for separating or decoding them. For example, whereas seizure and sleep waveforms are distinct and easily separable [1], EEG signals in BCI applications are typically subtle and require custom spatial and/or temporal filters [34]. This changes the discrimination techniques when dealing with BCI to spatial and temporal features [73, 125]. Auditory and visual stimulus response [2] have distinct spatial patterns as well adding to the diversity of BCI waveform morphology [75].

To distinguish spatial and temporal features, EEGs are partitioned via channels and epochs. As discussed previously, the channels are a representation of the electrodes, shaped by filtering and montages. Epochs segment the data as a function of time, typically on the order of seconds. Clinician and algorithm based approaches both rely on these techniques, but in different ways. Clinicians will review EEGs

using epochs on the order of tens of seconds [24, 126], while algorithms operate on epochs of seconds [14, 75].

One of the main diagnostic applications of EEGs is the classification of seizures [14]. Seizures represent excessive electrical activity within a region of the brain which manifest as high energy waveforms. The study of sleep is also an active research area given the occurrence of seizures during sleep and sleep's impact on brain health [102]. When recording for seizure and sleep activity a substantial amount of background activity is also captured. This enables enables an analysis of overall brain function, like the presence of ADHD in children[127]. Adult EEGs also provide insight into numerous conditions such as alcoholism [46], Alzheimer's Disease [78], brain development [128], emotion [47], and stress [80].

In a research setting, BCIs promote a deeper understanding of brain functionality by allowing those with disabilities to communicate [2] and regain functionality [70]. BCIs highlight the ability of algorithms to classify waveforms beyond the capabilities of clinicians. These computer-driven methods enabled the development of novel applications in clinical monitoring, video games [5], and bio-metrics [129]. All of these use real-time classification which is not in the purview of clinicians. Specifically, biometrics provide the ability to dissect the facets of EEG that differentiate one person from another. This is a level of discrimination that clinicians cannot attain and serves needs far beyond clinical settings in hospitals.

Moving EEGs outside of hospitals has expanded the potential applications of EEGs[4]. It is easier to produce EEG datasets for experiments, but even with these advances there are few publicly available datasets. Those datasets available having varying levels of documentation and labeling related to conditions, subjects, and tasks. In addition, the sampling rates and number of channels have no definitive standards which furthers the disparate nature of the recordings. Recording in non-

clincial environments often increases the likeliehood of artifacts, but even under ideal clinical conditions artifacts are still present requiring pre-processing[87, 8].

The following sections focus on the process and techniques of collecting EEG signals from a brain. Electrode configuration and montages are two important tools clinicians use when making a diagnosis from a recording. They provide flexibility to the clinician, but hamper the ability of algorithms to validate themselves on similar data. The experimental datasets are also introduced to highlight the difficulties of working with publicly available data.

2.1.1 Properties of Electroencephalograms

An EEG is comprised of multiple surface/scalp electrode channels capturing the continuous signals generated by the brain. These signals represent the aggregated neuronal activity of the cortical neurons in immediate proximity to each electrode. Each channel maps to a specific electrode that is placed on the scalp, extracranially, or in the case of intracranial electroencephalograms (iEEGs) directly on the brain's surface. Electrode placement for extra-cranial recordings follows a standardized layout, figure 2.1, based upon relative distances [130]. Intra-cranial electrodes are high density electrode grids that are placed directly on the brain region of interest. This increases the complexity of the electrode and the data collected which excludes them from this work, but there is no theoretical reason I-Vectors could not operate on such signals.

The electrode configuration dictates the number of channels in the recording. To visual these signals clinicians view them indirectly as *montages*, a differential electrode configuration. Montages, table 2.1, can be configured to be referential to a common ground electrode, neighboring electrode, or a contralateral electrode. These configurations aid in the diagnostic process by calling attention to patterns

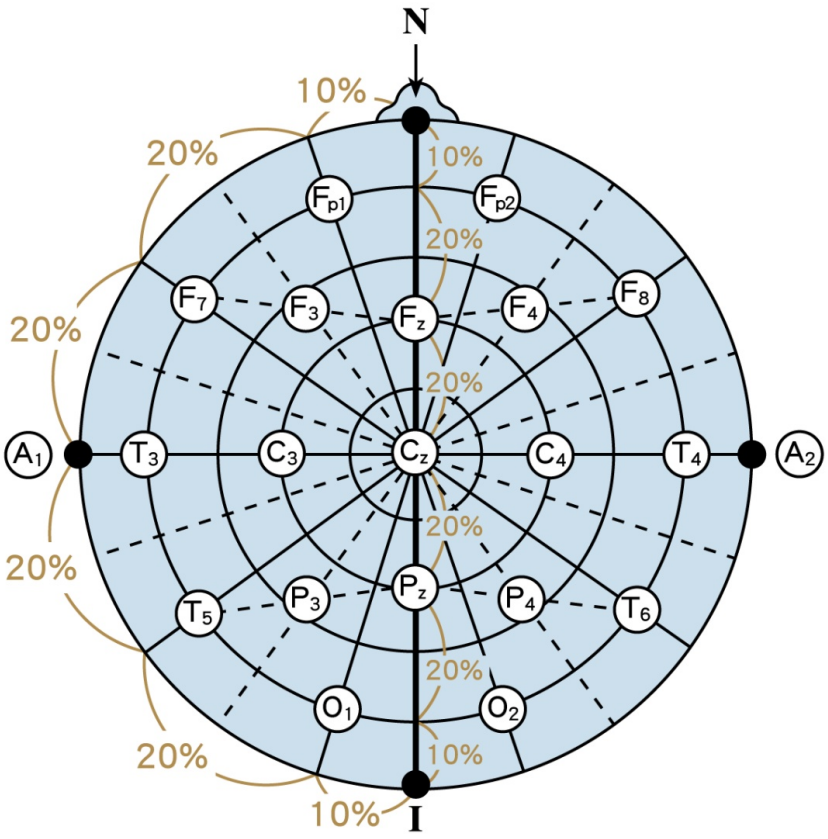


Figure 2.1: The 10-20, 10-10, and 10-5 layouts for EEG electrodes utilize a proportional unit of measure for the distribution of electrodes. The first number represents the distance of the electrodes from the nasion and inion and the second represents the space between subsequent electrodes. With this approach adding electrodes does not change the location of the previous electrodes. Image sourced from [131].

of behavior in the recording. Below are three sets of montages for a system with eighteen channels[1].

Montages serve to improve the clarity of each channel. Theoretically they do not impact the content of the channels, but evaluating such a claim is beyond the immediate focus of this work. Filtering of the channel data, before or after inclusion in a montage, is necessary to separate signals into the five standard EEG frequency bands, table 2.2. Signals between 2Hz to 80Hz represent the spectrum commonly

[1]Taken from: `https://www.acns.org/UserFiles/file/EEGGuideline3Montage.pdf`

Table 2.1: Table of EEG Montages

Channel	Longitudinal Bipolar	Transverse Bipolar	Referential to Ground(Ear)
1	Fp1-F7	F7-Fp1	F7-A1
2	F7-T3	Fp1-Fp2	T3-A1
3	T3-T5	Fp2-F8	T5-A1
4	T5-O1	F7-F3	Fp1-A1
5	Fp1-F3	F3-Fz	F3-A1
6	F3-C3	Fz-F4	C3-A1
7	C3-P3	F4-F8	P3-A1
8	P3-O1	T3-C3	O1-A1
9	Fz-Cz	C3-Cz	Fz-A1
10	Cz-Pz	Cz-C4	Pz-A2
11	Fp2-F4	C4-T4	Fp2-A2
12	F4-C4	T5-P3	F4-A2
13	C4-P4	P3-Pz	C4-A2
14	P4-O2	Pz-P4	P4-A2
15	Fp2-F8	P4-T6	O2-A2
16	F8-T4	T5-O1	F8-A2
17	T4-T6	O1-O2	T4-A2
18	T6-O2	O2-T6	T6-A2

viewed by clinicians[2]. For many conditions the frequency range of activity is critical in signal classification. Motor activity signals dominate the alpha band [132], while the stages of sleep affect all but the gamma band [35].

2.1.2 Available Datasets

There are a number of publicly available EEG datasets [3]. These datasets are developed for specific studies independently of each other resulting in a wide variation of data content and format. Their data formats range across European Data For-

[2]While this is the dominant spectrum of interest, research using iEEGs indicates activity at higher frequencies (>500Hz) may contain relevant discriminatory data related to seizures [109].

[3]The University of California San Diego maintains a website, `https://sccn.ucsd.edu/~arno/fam2data/publicly_available_EEG_data.html`, indexing many of the publicly available datasets.

Table 2.2: Table of EEG Frequency Bands. *When dealing with motor cortex signals it is common to encounter the Mu band (9-11Hz) which resides within the Alpha band.

Band Name	Frequency Range (Hz)	Attributes
Delta	1-3	Brain health, deep sleep
Theta	4-7	ADHD rhythms, relaxation
Alpha*	8-12	motor activity, alertness
Beta	13-30	anxiety, focus
Gamma	31-80	REM sleep, stress

mat (EDF), Matlab formatted files, and raw text files. The data content differs in terms of electrodes, sampling rates, and the studied phenomena.

This work applies to the PhysioNet Database dataset and the Temple University EEG Corpus (TUH Corpus) dataset. These datasets have been standardized to utilize the same 20 channel Trans-Cranial Parasagittal (TCP) montage. In addition the TUH Corpus dataset contains annotations from multiple sources providing robust labeling of events. This helps control for variation between the BCI focused PhysioNet Database dataset and predominantly seizure focused TUH Corpus dataset.

2.1.2.1 Temple University Hospital EEG Corpus

The TUH Corpus dataset contains over 25,000 EEG studies and their associated neurological evaluations taken from Temple University Hopsital (TUH) in Philadelphia, Pennsylvania [18]. Each patient's records present with different electrode configurations and sampling rates. The curated corpus uses a common 22 channel montage, TCP shown in figure 2.2, for all subjects with a static sample rate of 250Hz.

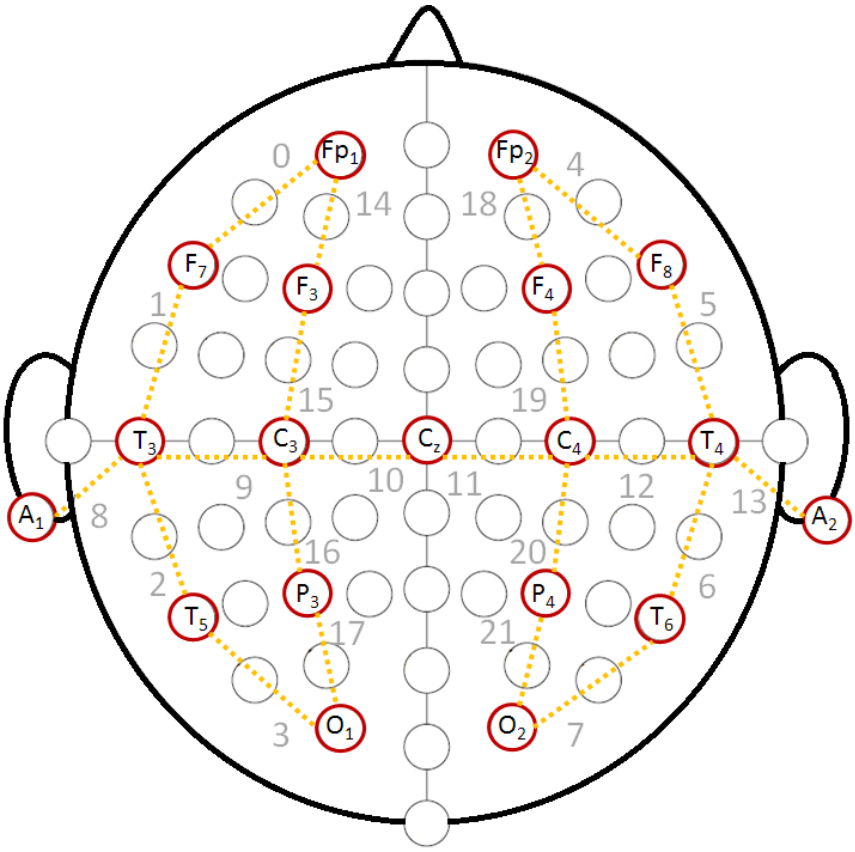


Figure 2.2: The TCP Montage channels (red) used by the TUH EEG corpus is overlaid on the PhysioNet Database channel layout. Each montage link (orange) is assigned an index for storing the montage channel (gray) data in the corpus. The proper 10-20 channel names (black) are provided for the montage channels.

The dataset contains longitudinal results of patients receiving continuing care at the hospital. These include multiple same patient sessions in a given day or sessions spaced out over a number of years. TUH treats patients of varying backgrounds (age, gender, diagnosis) providing breadth to the data. Recording profiles at TUH range from 23 to 32 electrodes with sampling rates of 250Hz, 256Hz, 400Hz, or 512Hz [18]. Computerized EEG analysis is complicated by the fact that even small variations in electrode placement can hamper generalizations between subjects. This problem is exacerbated when datasets from disparate sources are combined.

2.1.2.2 PhysioNet EEG Motor Movement/Imagery Database

The PhysioNet Database data contains 109 subjects following computer prompted motion/motion imagery trials at the New York State Department of Health's Wadsworth Center [94]. The recordings present 64 electrodes following a 10-20 layout sampled at 160Hz. From this base layout, the data is converted to the same 22 channel TCP montage used by the TUH Corpus.

Each subject performs two calibration trials (resting eyes open and resting eyes closed) and twelve task driven trials. The four tasks consist of opening/clenching the (1) left or (2) right first and opening/clenching both (3) fists or (4) feet as a physical and imaginary movement. A trial consists of 30 tasks that alternates between rest and motor tasks. The calibration trials last for one minute and the motor trials last for two minutes, providing 26 total minutes of subject data. The data is publicly available through the PhysioNet Database website [100].

There are 12 total motion tasks representing three groups. These groups consist of 4 repeated trials creating natural cohorts of grouped trials: {3, 7, 11; 4, 8, 12; 5, 9, 13; 6, 10, 14}. figure 2.3 shows the layout of tasks within each trial and their associated grouping. The major experiments utilize these trial level cohorts and the unique 109 subjects to develop I-Vectors for discrimination on the trial and subject level.

2.2 Applications and Classification of Electroencephalograms

The techniques used by algorithms and clinicians to classify and cluster EEG data are unique. An algorithm's foundation is informed by the knowledge of clinicians via their annotated data. A clinician's knowledge comes from their experience treating patients and their formal education. The algorithms are dependent on the clinicians'

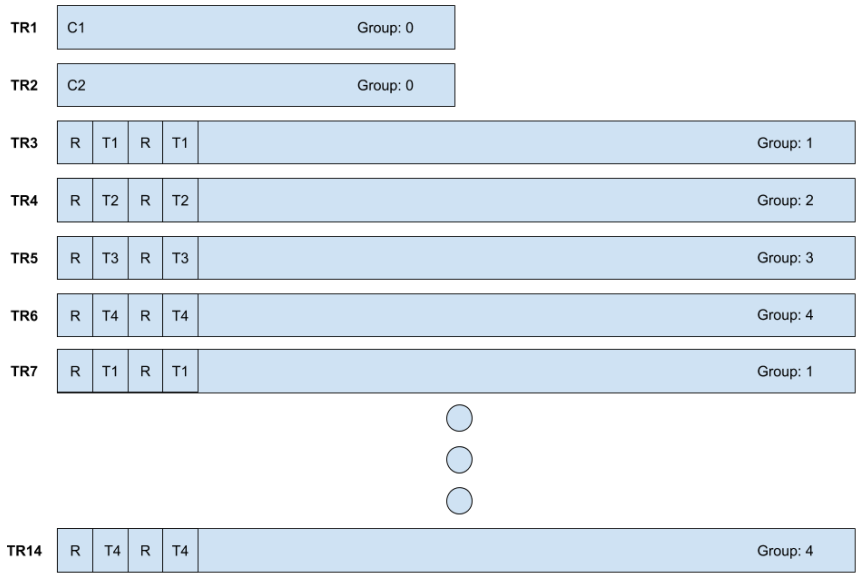


Figure 2.3: Each subject from the PhysioNet data set completed 14 trials. Two of these trials (TR1 and TR2) are one minute calibrations trials of resting eyes open and resting eyes closed. The remaining 12 trials are two minute recordings of a predefined sequence consisting of a task state and resting state. With four tasks states, each task is repeated three times producing four groups of task related trials. These trial groups provide the basis for cohort retrieval on the trial level.

annotations to build their knowledge base, making them susceptible to clinician bias. Clinicians are skeptical of algorithm performance because it does not match clinical performance. As algorithms attempt to improve their classification they are competing against experts in a field that is still being understood. Progress is slow because it is difficult for algorithms and clinicians to be confidant in the reasoning of their classifications. This makes it difficult to produce accurate testing datasets given the competing views on what are accurate annotations.

Clinicians annotate EEGs recordings to diagnose their patient. Typical clinical recordings are 20 minutes or more depending on the nature of the assessment. Each recording is accompanied by a detailed EEG report [27]. These reports must docu-

ment the subject, the testing carried out, and address the *clinical questions*[4]. The interpretation of an EEG recording is the main criteria when affirming a diagnosis, but must be supported by evidence indicating the recording is normal or abnormal[19].

This annotation and reporting process relies on the clinician's ability to review segments of the full recording for waveforms relevant to the clinical questions. A clinically relevant interpretation of the patient's condition may not be forthcoming without reviewing the reports of other tests and/or subjects [27]. This meta-analysis across subjects is a clustering process informed by medical records and annotations. However, the EEG reports focus on determining if the results inform the clinical questions or not [19]. This does not require all relevant phenomena to be annotated, as only enough data must be collected to affirm a position. As such a clinician's ability to cluster could be hampered by their ability to annotate, which is suggested by tracking a clinician's ability to reproduce classifications [26].

In contrast, an algorithm's approach to annotation is much more broad. Depending on the desired outcome, algorithms can perform a normal/abnormal classification [7], annotate specific epochs [14] or combine these approaches to classify EEG recordings [35]. Each of these classification techniques is a subset of the classification approach used by clinicians. Performance of these algorithms is measured against gold standards generated from training data annotated by clinicians [14, 24]. The goal is develop algorithms capable of mirroring clinical performance which limits the strength of the algorithms to the strength of the clinicians.

Depending on the output of these algorithms, they are capable of clustering EEG recordings in a way clinicians cannot replicate. The ability to infer similarity of wave-

[4]Clinical questions are posed prior to testing by the clinician. They serve to inform the clinician about the patient, their condition, and what outcomes are possible. As an example, if a patient has seizures while sleeping it would be necessary to determine the location of these seizures, their severity, and how such seizures compare to other patient populations. These would all be questions answered through EEG recordings.

forms, epochs, and entire recordings across subjects is important in the development of robust BCI [74] and bio-metric applications[45]. In this area algorithms exceed the ability of clinicians by shifting how EEG recordings are evaluated through novel channel and feature selection [65, 66, 68].

Specifically, bio-metric algorithms can determine the similarity of one subject to another [45, 63]. This makes bio-metric subject verification the closest analog to I-Vectors, but they are not limited to subject comparisons. Instead they offer the ability to discriminate on multiple facets of the data without needing the same extent of bio-metric pre-processing [121]. This makes their application to EEG recordings interesting as I-Vectors may be capable of bridging classification between algorithms and clinicians.

2.2.1 Clinician Classification

For clinically annotated EEG recordings it is important that common terminology was used when describing the waveforms. Without a shared vocabulary EEG reports would be ineffectual for diagnostics and documentation[27]. Gaspard et al.[22] tested 49 clinicians' agreement on terminology by asking them 409 questions about 37 pre-selected EEG waveforms. Their protocol removed the need of the clinician to find the epochs, enabling them to focus on each clinician's ability to describe the contents of each pre-selected epoch.

Each clinician's background varied in terms of experience (2-15+ years) and training (adult or pediatric neurology). While the epochs were sourced from only critical care patients exhibiting PLEDs, GPEDs, seizures, and other rhythmic activity. The epochs were presented using a modified biploar montage with a bandpass filter spanning 1Hz-70Hz. From these epochs, clinicians made *categorical assessments* based upon the presence of a seizure and dominant morphologies and *ordinal assessments*

based upon the physical properties on the signals (sharpness, amplitude, frequency, etc). The overall and inter-rater agreement of the clinicians is presented in table 2.3.

Table 2.3: Each terminology item, aside from Seizure, could be classified with multiple responses. Fast Activity could be yes, no, or no applicable while Phases were 1, 2, 3, >3, not applicable forcing the clinicians to articulate their classifications. Agreement specifies the percentage of waveforms classified correctly. The κ score indicates the amount of inter-rater agreement, see **??**.

Terminology Item	Agreement (%)	κ statistic (95% CI)
Categorical		
Seizure	93.3	91.1 (90.6-91.6)
Main Term 1	91.3	89.3 (89.1-89.6)
Main Term 2	85.2	80.3 (79.4-81.2)
Triphasic Morphology	72.9	58.2 (56.1-60.2)
Plus + Modifier	49.6	33.7 (32.4-35.1)
Any +	59.3	19.2 (17.5-20.9)
+ Fast Activity	71.9	65.5 (64.4-66.7)
+ Rhythmic Activity	76.5	67.4 (66.5-68.3)
+ Spike or Sharply Contoured	83.9	81.8 (81.2-82.5)
Ordinal		
Sharpness	91.5	84.8 (84.3-85.2)
Absolute Amplitude	96.5	94.0 (93.8-94.2)
Relative Amplitude	71.8	66.4 (65.3-67.4)
Frequency	97.8	95.1 (94.9-95.2)
Phases	89.9	83.0 (82.6-83.4)
Evolution	65.6	21.0 (19.7-22.2)

In 12 of the 15 categories, the clinicians' exceeded an agreement of 70% and 7 of the 15 showed near- perfect (0.81-1.00) κ statistics. The categories with the lowest agreement and weakest κ statistics were categorical classifications. With only 3 morphologies reporting κ below substantial (0.61-0.80), the results suggest the clinicians perform well as a group. Yet, those three categories indicated a universal blind spot that would be passed on to an algorithm built from this annotated data. Since the

contents of epochs were known, this showed how difficult it was for clinicians to agree on labeling of wavefroms.

These biases likely existed because clinicians were evaluated on their annotations indirectly. Their diagnoses were not solely based on a single event in the EEG, but rather the sum of the recordings in conjunction with the patient's medical history. In Halford et al. [26] the importance of detecting epileptiform transients (ETs) was found to be critical for diagnosing epilepsy. Failing to annotate some of the ETs does not change the diagnosis because the clinicians were primed to make a decision about epilepsy. Individually the 18 tested clinicians were unable to produce a Gwet agreement coefficient[5] over 0.50 with the rest of the group. This indicated a weak agreement among the clinicians. Despite varying levels of certification and years of practice, there were no distinct indicators of what characteristics represented a better annotator.

The difficulty in producing accurate annotations with respect to others existed at the intersection of finding the waveforms and then correctly labeling them. These problems were documented to various degrees when clinicians' annotation skills were tested on critically ill patients [31], patients exhibiting seizures [21, 23], comatose cardiac patients [29], and sleeping subjects [126]. The results of such studies highlighted problems with clinician inter-rater and intra-rater agreement as a function of the type of EEG data.

2.2.1.1 Clinician Inter-rater Agreement

The previous section discussed this broadly and with the benefit of the waveforms being pre-selected. However, when clinicians were asked to annotate longer epochs

[5]The Gwet's AC2 is an alternative to κ statistics for quantifying inter-rater similarity, but is bounded over the same range [133].

the discrepancies shift from clinical knowledge to issues of annotation style. Their inter-rater agreement was the ability of one clinician's classification to agree with one or more other clinicians.

A pedantic instance of this was seen in figure 2.4 where two clinicians labeled seizure events [23]. In the highlighted section, Rater B identified two discrete events while Rater A labels them as one event. Each of them notices at least 4 other seizure events, but their agreement was weakened because of their three misidentified events. Behavior such as this further complicated how to quantify agreement and disagree based upon duration of said annotations.

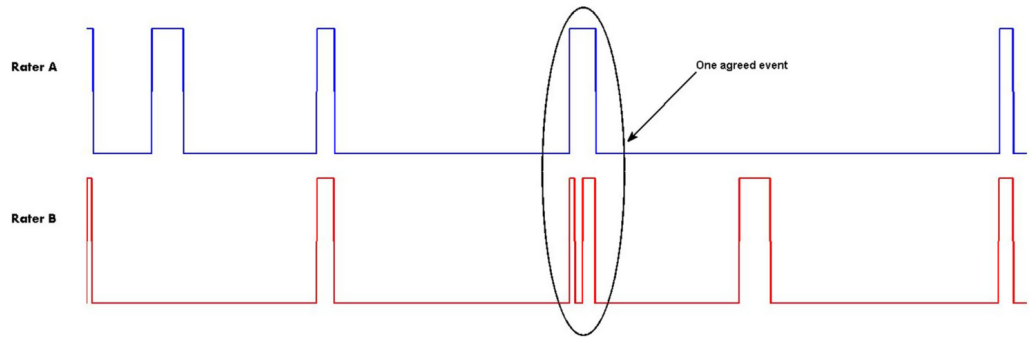


Figure 2.4: An example of how open ended annotation styles lead to inconsistencies in evaluating the accuracy of inter-rater agreements.

This example came from Halford et al.[23] where the agreement of 8 clinicians was test on 30 one hour Intensive Care Unit (ICU) EEG recordings from 20 seizure patients. Each clinician was asked to label PDs events, a strong indicator of a seizure, and true seizure events. The resultant κ statistics for the group were 0.58, moderate, for seizures and 0.38, fair, for PD. These results highlighted the difficulty in finding consensus by suggesting it surpassed their background and experience. There was a clear issue in how clinicians selected waveforms in the recordings, which resulted in less data being included in any gold standard.

Gerber et al.[31] conducted a study with a more expansive classification list than Halford et al.'s by expanding the available labels and varying the amount of available data. Two data sets, split into epochs of 10 seconds and epochs >20 minutes, were built from 11 subjects with convulsive seizures, status epilepticus[6]. The results, table 2.4, showed the clinicians' consensus was stronger on the shorter epochs (0.04-0.68) than the longer epochs (0.07-0.44).

Table 2.4: Results of classification using segments of 10 seconds and > 20 minutes in length. Five clinicians annotated the shorter epochs and all seven clinicians annotated the longer epochs. The κ statistics for both datasets are reported along with the raw agreement percent for the 20min epoch dataset.

Term	10s Epoch Kappa	20min Epoch Kappa	20min Epoch Agreement (%)
Rhythmic/periodic vs. excluded	0.68	0.44	82
Localization	0.49	0.42	66
Morphology	0.39	0.37	69
Frequency	0.34	0.27	78
"Quasi" vs. Not	0.04	0.07	57
"Frontally Predominant" vs. Not	0.40	0.08	68
+ vs. Not	0.12	0.08	62

The most critical labels (rhythmic/periodic vs. excluded, localization, and morphology) exceed 65% agreement, but only rhythmic/periodic exceeds 80%. This meant that on average each clinician failed to recognize 20% to 35% of what the other clinicians annotated. Without definitively labeled data it was impossible to determine if the 35% gap is due to false positives or false negatives. Such knowledge could be used to determine if they were over-jealous or overly-shrewd in their annotations. However, it was possible their performance was impeded by alignment issues similar to those seen in Halford et al.'s work. The results otherwise suggested that

[6]Status epilepticus is the categorization of a person's state when seizures occur close together or occur for a prolonged duration(>5 minutes).

the clinicians agree at a moderate to fair level which was enough to make accurate medical decisions, but not sufficient from which to train algorithms.

Gerber et al.'s best reported inter-rater agreement was inline with Halford et al.'s. This trend persisted in the work of Grant et al.'s work [21]. Their study evaluated the agreement of 6 clinicians (adult and pediatric neurologists) classifying 7 categories (status epilepticus, seizure, epileptiform discharges w/ and w/o slowing, slowing, normal, uninterpretable) of waveforms in 150 30 minute EEG epochs. Each clinician reviewed a unique set of 150 epochs from the full dataset's 300 30-minute epochs. Over the 15 inter-rater pairs, their inter-rater κ scores ranged from 0.29 to 0.62 suggesting fair to substantial agreement among the pairs.

Table 2.5: Inter-rater agreement for the 15 clinician pairs observed by Grant. The pair averaged κ score is 0.44 giving the overall agreement as moderate.

Reader Pair	κ score
AB	0.43
AC	0.52
AD	0.37
AE	0.37
AF	0.50
BC	0.48
BD	0.41
BE	0.37
BF	0.29
CD	0.49
CE	0.56
CF	0.62
DE	0.48
DF	0.35
EF	0.42

Westhall et al. [29] had a smaller subject pool, 4 clinicians, but asked them to evaluate EEG recordings for specific to *Prespecified EEG patterns*, *Background EEG*, or *Periodic or rhythmic patterns*. Each $>$ 20 minute recording was drawn from a

pool of 103 comatose cardiac arrest patients. For the prespecified EEG patterns the κ statistics ranged from 0.42 to 0.71, table 2.6. Meanwhile, the background and periodic patterns produced inter-rater κ statistics between -0.07 to 0.82, table 2.7.

Table 2.6: Agreement and Kappa statistics using the ACNS classification labels for inter-rater performance on prespecified EEG patterns.

EEG Waveform	Agreement (%)	κ statistic
Highly Malignant	75	0.71 (0.55-0.79)
Malignant	63	0.42 (0.34-0.51)
Benign	63	0.42 (0.34-0.51)

Just as the results of Gerber et al. showed strongest performance for critical waveforms, Westhall et al. did too. However, performance outside these critical waveforms was extremely poor in terms of classification agreement and κ statistics. This might have been caused by the increase in classification categories, compared to Gerber et al., Grant et al., or Halford et al, but more likely suggested the clinicians fundamentally disagreed over the non-prespecified EEG patterns given their previously discussed terminology consensus. Conversely, if background EEG or periodic patterns were necessary to make a diagnosis it would be difficult to resolve an understanding from the work of these clinicians.

2.2.1.2 Clinician Intra-rater Agreement

Clinicians difficulty in producing acceptable κ statistics in inter-rater testing extended themselves via intra-rater testing as well. In most cases, intra-rater agreement addressed a clinician's ability to reproduce annotations on data they previously annotated. Gerber et al., Grant et al., and Westhall et al. ran specific intra-rater experiments to track inter-rater behavior.

Table 2.7: A breakdown of the ability of clinicians to adequately annotate background events and repeated EEG patterns.

	Inter-rater Agreement (%)	κ	Intra-rater Agreement (%)	κ
Background EEG				
Continuity	37	0.76	62	0.86
Voltage	47	0.65	75	0.31
Predominant Frequency	3	0.36	30	0.17
Reactivity to sound	42	0.25	82	0.76
Reactivity to pain	32	0.17	69	0.44
Periodic or rhythmic patterns				
Periodic or rhythmic discharges	50	0.56	80	0.55
Prevalence	39	0.49	70	0.58
Typical frequency	6	0.82	55	0.80
Maximum frequency	14	0.74	54	0.68
Sharpness	74	0.73	75	0.58
Absolute amplitude	44	0.42	86	0.59
Stimulus induced pattern	63	0.19	80	0.48
Evolution	13	0.19	76	0.30
Plus Modifier present	19	0.17	84	0.28
Triphasic morphology	61	-0.07	63	0.00

Gerber et al. evaluated the ability of 5 clinicians to reproduce their results on the 10 second epochs 12 months after the original study. The same epochs were used, presented in a randomized order, and each clinician was asked to follow the classification scheme from the original study. The resultant κ statistics, table 2.8, showed the difficulty clinicians had in agreeing with themselves. Compared against inter-rater agreement, table 2.4, the intra-rater agreement was only marginally better.

The follow-on experiment in Grant occurred 4 months after the initial study. In this case, the range of intra-rater agreement (0.33 to 0.73) was better than that of the inter-rater agreement (0.29 to 0.62). However, the intra-rater results suggested clinician A was the worst performer. This conflicted with clinician A's inter-rater

Table 2.8: The 5 clinicians in the original 10s epoch evaluations, re-evaluate the same set of data 12 months later. These results represent how well each clinician agrees with their original classifications.

Clinician	Rhythmic/ Periodic vs. Excluded	Local.	Morp.	Freq.	"Quasi" vs. Not	"Frontally Predominant" vs. Not	"Plus" vs. Not
1	0.79	0.58	0.67	0.30	0.28	0.32	-0.03
2	0.86	0.60	0.55	0.24	0.25	0.38	0.00
3	0.68	0.51	0.15	0.28	0.32	0.45	0.28
4	0.73	0.68	0.58	0.29	-0.08	0.57	0.24
5	0.76	0.46	0.40	0.19	0.28	0.67	0.00
Mean κ	0.76	0.57	0.47	0.26	0.21	0.48	0.098

agreements, table 2.5. The worst inter-rater agreements did not involve clinician A, but rather clinicians B, D, and F. These results suggested inter- and intra-rater agreement scores were poor tools for understanding a clinician's annotation ability, but confirmed their ability to generate consistent diagnoses.

Table 2.9: The 6 clinicians were tested twice 4 months apart. These agreement scores represent their intra-rater consensus on 7 classification categories.

Clinician	κ score
A	0.33
B	0.50
C	0.58
D	0.67
E	0.73
F	0.64
Mean	0.59

The trend of intra-rater agreement, table 2.10, scoring higher than inter-rater agreement, table 2.6, was repeated by the clinicians Westhall et al tested as well. Repeating their original experimental protocol 6 months later produced very high intra-rater classification agreements, table 2.10. However, the κ statistic for highly malignant, 0.64, was lower than its inter-rater counterpart, 0.71. Despite each clin-

ician improving and/or maintaining their classification ability, they were unable to identify the same waveforms as they did previously. This again spoke to nature of clinicians ability to only need a minimum amount of insight to generate a consistent diagnosis.

Table 2.10: Agreement and Kappa statistics using the ACNS classification labels for intra-rater performance.

EEG Waveform	Agreement (%)	κ score
Highly Malignant	88	0.64 (0.48-0.83)
Malignant	98	0.93 (0.57-1.00)
Benign	98	0.93 (0.57-1.00)

The other features in table 2.7 represented less discrete facets of EEG waveforms. These features required qualitative analysis which increased the difficulty of classification consensus, exemplified by the abundance of slight and poor inter-rater κ statistics. Intra-rater agreement showed minimal improvement of κ statistics, while the averaged intra-rater agreement % was better than its counterpart. This suggested clinicians were capable of reproducing their work, but were prevented from doing so by their innate biases thus limiting their κ statistics.

As a whole these intra- and inter-rater studies indicated clinicians were consistent within themselves, and their cohorts, when classifying EEG recordings. However that consistency did not appear to translate into producing data acceptable for use as a gold standard. While the results of each study offered suggestions as to why such consensus was difficult to reach, there was no single conclusive factor. The size of the epochs, the category of classification, the duration of the annotated waveform, and the clinician's training and experience all impacted the resultant κ statistics. Their inability to come to agreement did not, however, diminish their ability to diagnosis. The only shortcoming was that it limited the quality and quantity of data available on which to train ML algorithms.

2.2.2 Algorithm Classification

Despite robust waveform nomenclature, translating EEG signals into features for algorithm classification was an open field. With no feature consistency, each study was free to develop their own features such as using a unique feature set [14], borrowing from a previous study's features [134], or forgoing features and using the raw data directly [135]. Regardless of the type of features, they all segmented the recordings into *epochs* which served as the input to the algorithms.

Most epochs represented a window in time, typically on the order of seconds, that contained the data from one or more EEG channels. The duration of the epochs drove a trade off between categorizing phenomena occurring rapidly, PDs, or slowly, such as sleep states. Given the number of channels in a recording, their duration, and the sampling rate EEG recordings typically produced significant amounts of data. The use of epochs was the first step of dimensionality reduction by attempting to normalize the raw data into manageable segments across channels, subjects, sessions, and datasets.

Thus the features used for these epochs needed to excel at minimizing the amount of data while maximizing the information density relative to the data type. This was a difficult task given the depth of EEG signals which was why feature sets were frequently developed for specific use cases like seizures [13], BCIs [136], sleep [24], alcoholism [46], ADHD [42], and beyond. The combinations of features and epochs allowed each study to focus on their specific goals, but made it difficult to produce a robust universal feature set.

This problem was compounded by the EEG community's continual adaption of the newest ML algorithms in an effort to increase classification performance. This behavior was not much different from the development of speech technologies until

they resolved a robust universal feature set [137] as they developed a myriad of techniques to address their classification problems, such as K-Nearest Neighbors (KNNs), Support Vector Machines (SVMs), Neural Networks (NNs), and GMMs. Often a given a combination of features and datasets performed better or worse than another depending on the algorithm and its parameters. This made it hard to determine if performance gains were due to algorithms, dataset, feature set, or something else.

The following sections reviewed algorithms that used *statistical models*, *supervised algorithms*, and *unsupervised algorithms* common to the EEG classification landscape. Statistical models formed the basis of numerous ML techniques and were frequently used to filter out artifacts via thresholding, detect ERPs, or interpret common spatial patterns (CSPs). Supervised algorithms used labeled data from clinicians and *a prior* knowledge to build classifiers focused on specific phenomena like seizures and mental states. Meanwhile, unsupervised algorithms leveraged the power of statistical models built from large unlabled datasets to classify conditions for which annotations were hard to obtain. These techniques were applied at one time or another on datasets generated from sleep, seizures, ADHD, or BCI EEGs.

2.2.2.1 Statistical Algorithms

Statistical modeling of known EEG phenomena provided a robust platform for developing basic classification algorithms. The type of modeling depended on the waveform, similar to how features were adapted, but classification was primarily based on one-versus-all evaluation. These approaches were mathematically straightforward and required minimal data relative to the defined phenomena. Their success, however, was data dependent as they required a thorough set of labeled data to operate. This made them ultimately reliant on the knowledge on clinicians.

An ERP represented an involuntary response by the brain when it perceived a targeted external stimulus. One of the most common instances of these events, the P300 response, was used to development basic BCI spellers. A P300-spellers were built to detect responses to auditory and/or visual stimulus enabling a person to spell words with their brain [76]. This phenomena was ideal for statistical modeling as brief subject specific training readily produced acceptable performances [138].

Guger et al. [138] showed that 5 minutes of training were enough to elevate the majority of the subjects to 60% or better accuracy, table 2.11. The training period asked the subjects to spell specific words and then used Linear Discriminate Analysis (LDA) to tune the weights of the 8 pre-selected channels. Subjects operated the speller by responding to a single character being flashed, single character speller, or by alternating flashing of rows and columns, row-column speller.

Table 2.11: ERP based spelling performance as a function of method.

Classification accuracy (%)	Row-column speller % of sessions 81 subjects	Single character speller % of sessions 38 subjects
100	72.8	55.3
80-100	88.9	76.3
60-79	6.2	10.6
40-59	3.7	7.9
20-39	0.0	2.6
0-19	1.2	2.6

This approach represented a highly effective real-time communication platform that did not require excessive training data nor overly complex signal processing. The main drawback was the time required to produce a single letter, 28.8 seconds for row-column spelling and 54 seconds for single character spelling. The technique itself was very specific to ERPs which meant it did not contribute much to other EEG applications. This necessitated the development of different statistical models for

addressing the detection of Alzheimer's Disease (AD) [90], ADHD [139], and seizures [140] events.

The ability to detect and classify seizures has remained a core focus of EEG research in terms of reviewing existing recordings as well as enable accurate predictions. Chu et al. [13] applied *attractor states*[7] to EEG data in an effort to improve seizure prediction and detection via statistical discrimination. The technique was tested on two datasets, the Children's Hopsital of Boston Massachusettes Institute of Technology Scalp EEG Database (CHB) and adult seizures from the Department of Neurosurgery of Seoul National University Hospital, using 50% overlapping channel independent 20s epochs. The raw epochs were converted to frequency banded Fourier coefficient features used to build seizure and non-seizure state models.

Their seizure predictions, using a 30 second horizon, averaged 90.20% sensitivity on the training data and 86.67% sensitivity on the testing data (2 subjects reported 0%). Decreases in sensitivity correlated with a drop in average false positives per hour from 0.476 on the training data to 0.367 on the testing data. The peak rate of false positives were 1.667 and sensitivity for multiple subjects was 0.0%. The results suggested a simple model can predict seizure onset, correctly predicting 39 of the 45 documented seizures across the 17 subjects. However, the failure to detect anything for 2 subjects (1 seizure each) and missing 2 seizures from another subjects indicated the technique may not be sufficient for all types of seizures nor all patients.

Understanding sleep cycles aided in understand seizures given seizures frequently occur at night [6], but first the stages of sleep needed to be classified. Warby et al.[24] compared the performance of six statistical sleep spindle, sleep stage markers,

[7]Attractor states are stable states which the data trends towards given its natural behavior. The concept originated from the work of Scheffer et al.[140], but is beyond the scope of discussion in this work.

algorithms[8] against clinicians and non-experts. The dataset consisted of 32,112 25s single channel epochs from 110 healthy subjects split into training, testing, and verification data. The verification data, built from 2,000 epochs scored by 5.3 clinicians on average, serves as the gold standard.

Each of the algorithms applied different flavors of energy thresholding (Root Mean Squared (RMS), PSD, or Fast Fourier Transform (FFT)) on a bandwidth filtered (9-16Hz) portion of the epochs. The algorithms' performances, table 2.12, were not in agreement with the gold standard (GS), but they did agree with the automated group consensus (AGC). Overall, the algorithms were the weakest classifiers while the clinicians were the strongest at classifying sleep spindles. The non-experts performed better than the algorithms which suggested these statistical based algorithms may not be an effective classification technique for this task.

Table 2.12: The sleep spindle detection agreement, evaluated as F_1 scores, shows the relationship between each algorithm and the expert group gold standard (GS), non-expert group consensus (NGC) and automated group consensus.

Algorithm	GS	NGC	AGC
a1	0.28	0.22	0.28
a2	0.28	0.30	0.40
a3	0.21	0.17	0.21
a4	0.50	0.46	0.79
a5	0.52	0.49	0.84
a6	0.41	0.37	0.48

Huang et al. [145] aimed to detect the presence of AD in a set of 93 subjects labeled as having AD, mild cognitive impairment (MCI) or healthy controls. Classification used the 15 2s epochs from each subject which were built on their alpha (8.0-11.5Hz) and theta (4.0-7.5Hz) global field potential (GFP), a generalized EEG

[8]The six algorithms were drawn from six unique studies cited here: {a1[141], a2[142], a3[143], a4[144], a5[99], and a6[57]}

Table 2.13: The table contains the mean values of the GFP for each frequency band over a given brain region. These represent the features the algorithms uses to discern AD subjects from MCI subjects and healthy controls.

Band	Group	GFP	Loc-X	Loc-Y	Loc-Z
Delta	AD	13.4(9.3)	12.5(9.8)	1.8(4.2)	-5.6(6.0)
	C	7.3(2.3)	12.9(8.6)	0.1(4.7)	-4.4(5.8)
	MCI	10.4(5.2)	12.2(11.3)	1.8(4.6)	-6.2(6.0)
Theta	AD	15.6(14.6)	-2.6(7.6)	2.1(5.2)	-0.2(6.9)
	C	8.0(6.5)	-5.7(7.2)	1.4(5.9)	-4.0(5.0)
	MCI	10.2(10.8)	-3.6(12.3)	2.7(5.5)	-2.0(6.3)
Alpha	AD	14.1(14.5)	-12.6(11.5)	-2.1(7.1)	1.7(8.9)
	C	31.2(30.2)	-21.0(7.3)	-0.4(5.4)	-3.4(7.2)
	MCI	40.1(43.3)	-19.9(11.1)	-0.1(6.3)	-1.7(9.3)
Beta 1	AD	3.7(3.7)	-6.2(11.2)	-1.5(8.9)	5.2(9.9)
	C	3.6(1.9)	-12.1(10.1)	2.2(5.8)	1.4(9.1)
	MCI	5.2(5.2)	-13.9(12.3)	1.3(6.9)	2.5(8.8)
Beta 2	AD	2.1(1.7)	0.3(12.8)	-2.3(10.8)	8.3(10.6)
	C	2.9(1.7)	-8.2(11.8)	1.8(7.4)	4.4(8.6)
	MCI	4.2(4.6)	-8.8(13.9)	1.0(10.4)	4.8(11.0)

amplitude measurement. The algorithm reported an AD classification accuracy of 84% against control subjects. This represented an optimal feature set, which started as epochs of FFTs decomposed into their GFP across frequency bands (delta (1-3.5Hz), theta, alpha, beta 1 (12-15.5Hz), and beta 2 (16-19.5Hz)). These features were then localized with respect to regions of the brain: antero-posterior (Loc-X), left-right (Loc-Y), and superior-inferior (Loc-Z). The resultant values of each feature permutation is shown in table 2.13.

ADHD was another omnipresent condition that could be detected through a subject's theta beta ratio (TBR) [146]. Lenartowicz et al. [146] reviewed multiple approaches for distinguishing ADHD patients from controls based on temporal and spatial features and the ratios of energy present in frequency bands and specific channels. The studies reported divergent performance when using TBR as a discrim-

ination metric. Monastra et al. [139] reported an accuracy of 91% (90% sensitivity, 94% specificity) while Buyck et al. [147] reported an accuracy of 49-55%.

Detecting ADHD through EEG recordings appears possible based on the TBR, but Lenartowicz et al. conclude the technique is not reliable enough to be a diagnostic test. The work of Monastra et al. was carried out in 2001, but advancement in the field, like Buyck et al.'s 2014 work, indicate variations in ADHD morphology make TBR a poor classification metric. Despite a clear clinical utility in using EEG recordings for ADHD diagnosis [148], the research suggested the condition was not yet understood to the point of being able to develop a robust statistical classification model for it.

However, Buyck et al. found that TBR did make an excellent, AUC 0.965, discriminator for age classification. This exemplified the difficulty in building a robust feature set for a given classification task as differnet sets of features could conflate multiple conditions. The best examples of this were the efforts made for detection and correction of EEG artifacts [8].

The most common artifacts (eye blink, muscle artifacts, and eye movements) were caused by the subject making them difficult to mitigate during recording. Jung et al. [149] indicated the overlap between artifacts and waveforms of interest prevents many novel artifact detection techniques from having a broader impact. Their work compared the performance of independent component analysis (ICA) to principal component analysis (PCA) on a dataset of normal and autistic subjects. Despite both techniques being capable of separating the signals from the noise, ICA offered the best performance for correcting the original recordings.

Delorme et al. [150] devised a more comprehensive experiment[9] for detecting artifacts. They applied six thresholding schemes to raw data and data processed each with ICA, building on the success of Jung et al. Their results, figure 2.5, showed that applying ICA improved the classification performance regardless of the artifact's source. However, the use of ICA did not improve the performance of each algorithm.

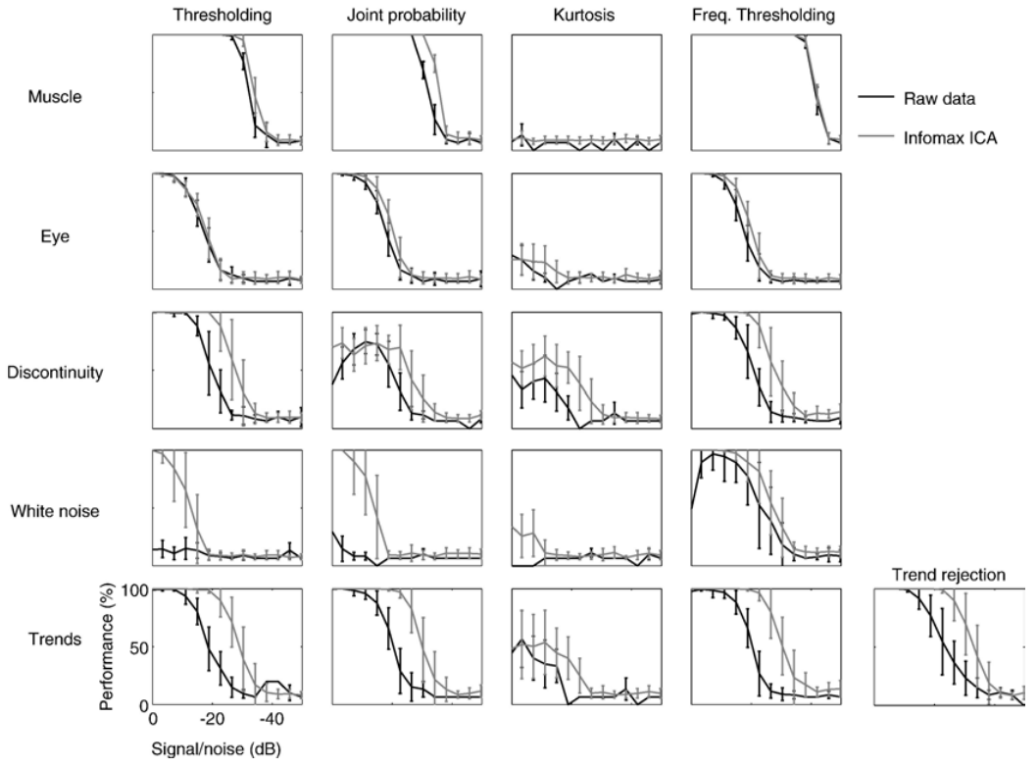


Figure 2.5: Classification performance of thresholding approaches based upon the signal to noise ratio of the artifact and the signal.

The largest changes in performance were related to the artifacts (discontinuity and white noise) and not the algorithms. This suggested that the algorithm's performance

[9]They compared five methods to determine how best to identify artifacts within a recording. (1) Extreme values: Artifacts detected if amplitudes exceeded a predetermined threshold. (2) Linear trends: Least squares thresholding against an average of the activity in an epoch. (3) Data improbability: Likelihood of an observations with respect to all observations from each channel. Each epoch became a product of likelihoods which should decrease if artifact events are detected. (4) Kurtosis: Measure the 'peakedness' of each epoch's distribution. (5) Spectral pattern: model scalp topology in conjunction with frequency spectrum.

does matter, i.e. Kurtosis performed universally poor, but ICA was only able to impact performance when the noise was distinct from the signal. This, of course, is the definition of ICA, but highlighted the problem of its use on EEG data where waveforms and artifacts presented as seemingly identical signals.

Despite this limitation, these experiments were mostly a success which lead to the development of Fully Automated Statistical Thresholding for EEG artifact Rejection (FASTER) [8] and Automatic EEG artifact Detection based on the Joint Use of Spatial and Temporal features (ADJUST) [40]. These techniques provided universal artifact detection and rejection across multiple types of EEG data. FASTER relied on a parameter set consisting of variance, Hurst exponent[10], amplitude range, and channel deviation over five thresholding levels (channel, epoch, epoch ICA, channel-epochs, and channel average). ADJUST used spatial and temporal feature extraction to classify and remove artifacts from ICA filtered data. These results again highlighted how different feature sets and algorithms achieved acceptable performance making it hard to know which option is 'right'.

Table 2.14: The sensitivity and specificity at the channel and epoch level for FASTER with respect to different channel configurations.

Channels	Channel Sensitivity(%)	Channel Specificity(%)	Epoch Sensitivity(%)	Epoch Specificity(%)
128	94.47	98.96	60.24	97.53
64	97.02	98.48	61.83	97.54
32	5.88	96.81	58.64	97.49

ADJUST was more complex, but achieved 60% sensitivity and 97% specificity, table 2.14, on 2 second channel independent epochs drawn from 47 subjects. FASTER was less complex than ADJUST and replicated the clinician's 95.2% artifact detection

[10]The Hurst exponent is a measure of the changes in lag observed from the auto-correlation of pairs of points in a time series.

performance. The most significant aspect of these results were that the testing dataset was built from 10 subjects that were withheld from the 21 subject training dataset.

Artifact detection and correction continues to be an active research topic, but the reliance on ICA remained. Mahajan et al.[87] reported exceptional performance using ICA on 12 electrodes followed by modified multiscale sample entropy (mMSE) and Kurtosis and thresholding. Their eye blink detection algorithm reported 90% sensitivity and 98% specificity across four subjects.

These statistical approaches to classification are promising, but they are developed on small datasets with simple goals. Adapting them for use on larger datasets with more extensive classifications needs seemed to be beyond their capability. At the very least they showed when EEG signals were broken down to their core components it was possible to reliably discriminate among them. This suggested reiterated the idea that it was possible ML algorithms to at least match a clinician's performance.

2.2.2.2 Supervised Algorithms

Supervised ML algorithms build statistical models from datasets with labeled classes. Each class would ideally represent a subset of related data (artifacts, sleep spindles, or ETs) that the algorithm would learn to distinguish between. Given a diverse feature set, the algorithms build decision surfaces based upon the strongest statistical properties of the features unique to each known class. These decision surfaces allowed classifications to be learned instead of having to infer them directly from the dataset.

These algorithms were setup with the aim of emulating a clinician's classification performance. In doing so, they tied themselves to the performance of those provided the labeled data. This is the main limitation of supervised learning: The algorithms must be shown what to classify making their success dependent on the properties of the training data. If the test contains a new class, the algorithm will struggle to define

it and it may go undetected unless additional analyses were undertaken. However, the strength of this approach is that supervised ML classification algorithms work extremely well for well known phenomena (artifacts, seizure, and sleep). This has been shown to be true even when such conditions occurred rarely or were learned from a small number of epochs [151]. This naturally meant they worked best paired with phenomena in smaller sets of clinically annotated data (BCIs, emotions, and workload) EEGs.

The classification of sleep relied on detecting waveforms known as k-complexes and sleep spindles which are unique to a sleeping brain. There is also generalized brain activity specific to the energy bands that accompany each stage of sleep[126]. Thus each stage of sleep contains a mixture of unique waveforms and shifts in the rhythms, ratios of energy in the EEG bands, and waveforms that make it distinct from other brain conditions. Such behavior is most notable in the that dominant (>50%) alpha rhythms where remain indicative of being awake. Stage 1 typically contains a split (50%\50%) of alpha and delta rhythms. Stage 2 contains sleep spindles and diminished (<20%) delta rhythms. Stage 3 sees a resurgence (20%-50%) of delta rhythms. Stage 4 and REM sleep are classified by dominant delta rhythms.

These discrete states made the adaptation of supervised ML algorithms straightforward. In Schluter et al.[35] the stages of sleep were classified with Decision Trees (DTs) by bagging[11] on an array of physiological data[12]. The classification was performed on 33,542 30 second epochs drawn from 15 subjects, table 2.15. On the whole separating wakefulness, REM sleep, and from the stages of sleep was excellent.

[11]Bagging, bootstrap aggregating, is a technique employed to reduce the variance of ML algorithms. The original data was re-sampled with replacement to produce multiple data sets containing redundant data.

[12]Sleep studies frequently collect electrocardiogram (ECG), EEG,electromyography (EMG), and electrooculography (EOG). In this work, aside from EEG data, EMG and EOG were used to help classify the sleep stages.

However, identifying the distinct stages of sleep proved difficult especially for stage 1 and stage 3. These results incorporated data in addition to the EEG recordings, suggesting EEG alone may not be sufficient for accurate classification.

Table 2.15: Confusion matrix of sleep stage classification covering wakefulness (W), each stage of non-REM sleep (S1,S2,S3,S4) and REM sleep.

	W	S1	S2	S3	S4	REM
W	97.0	2.4	0.6	0.1	0.0	0.5
S1	9.1	58.1	20.2	0.8	0.2	11.6
S2	0.5	4.7	91.7	5.5	0.8	0.2
S3	0.0	0.1	20.2	62.8	18.2	0.1
S4	0.1	0.2	1.0	12.6	86.8	0.1
REM	0.7	2.3	3.0	0.1	0.0	96.6

Radha et al. [102] used different algorithms to classifying the stages of sleep, but produced similar results to that of Schluter et al. Their data consisted of 30s epochs of 34 features drawn from 10 health subjects. They compared two supervised algorithms, Random Forest (RF) and SVM, ability to classify the epochs into REM sleep and the 3 stages of non-REM sleep (N1,N2,N3). By using supervised algorithms it was necessary to have a clinician provide labeled training data. However, this also allowed a κ statistic to be associated with each algorithm's performance relative to the clinician, table 2.16. Prior to classification the feature set was optimized for the differential montage channel (F4-A1), an epoch duration of 30s, and only 20 of the original 34 features.

These results were comparable, table 2.15, to Schluter et al., which was a study that used far more data. The moderate to substantial κ statistics suggested the algorithms performed as well as a clinician would have given the previously reported inter-rater agreements. However, it was possible that the feature optimization drove this performance. Sleep states were not a unique phenomena and they tended to represent major changes in brain activity, the necessity of channel and feature optimiza-

Table 2.16: Precision and recall of SVM and RF classification using a single EEG channel for sleep stage classification. In this study non-REM sleep is broken into only three stages (N1, N2, N3) making it difficult to compare to the standard four non-REM stages of sleep shown in table 2.15.

Sleep Stage	SVM 1vA Precision	SVM 1vA Recall	SVM 1v1 Precision	SVM 1v1 Recall	RF Precision	RF Recall
W	0.86	0.51	0.75	0.71	0.78	0.73
N1	0.00	0.00	0.18	0.00	0.52	0.31
N2	0.86	0.83	0.85	0.88	0.85	0.91
N3	0.32	0.70	0.82	0.70	0.83	0.73
REM	0.56	0.55	0.58	0.79	0.69	0.70
Accuracy	0.69		0.77		0.80	
κ	0.46		0.61		0.66	

tion suggests this algorithm/feature combination was only able to find the strongest indicator of sleep and may be missing out on the nuances of individual sleep stages.

Similar to sleep, seizures had frequently been categorized into distinct stages: *normal* indicative of a normal healthy state, *pre-ictal* indicative of a build up to a seizure, *ictal* indicative of an active seizure [152], and *post-ictal* indicative of the time following a seizure [13]. Accurate detection of these stages, specifically pre-ictal, could help improve the diagnosis and treatment of epilepsy [6]. Seizure classification was always a primary research focus of automated algorithms because of number of people affected by them[6]. Effort has been continually applied to improve the classification of seizures which tended to focused on developing better features than the FFT based frequency band powers [14, 54, 13] and improving algorithms [152, 25, 153]. These efforts were predicated on, and thus limited by, the availability of annotated data and the quality of the annotations.

Wulsin et al.[14] used raw data and diverse feature subsets derived from a stock listing of features [13] to compare seizure detection as a function of algorithms and features. Despite efforts to find a suitable feature subset, the strongest classification occurred when using the raw data as the input features. In addition to the feature analysis, four classification algorithms (DTs, SVMs, KNNs, and Deep Belief Networks (DBNs)) were evaluated with SVMs producing the best classifications, figure 2.6.

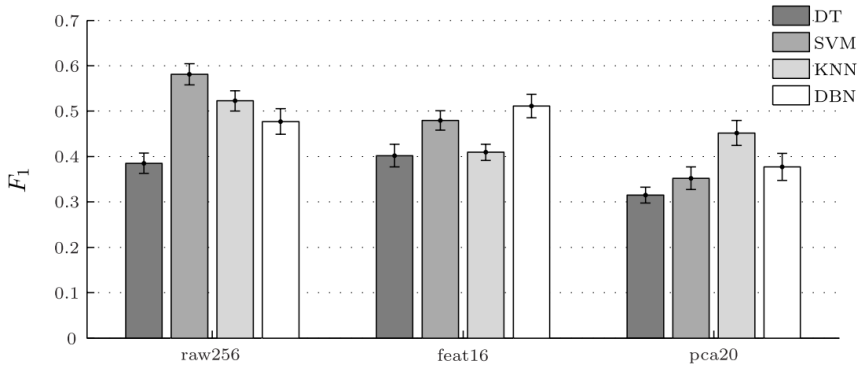


Figure 2.6: Wulsin et al. evaluate algorithm peformance based upon the F_1 measure, where $F_1 = 2 * (sensitivity * precision)/(sensitivity + precision)$. The results are presented to compare the algorithms and feature sets against each other. The feature sets are comprised of: *raw256* represents the raw waveform data, *feat16* are the hand selected 16 features, and *pca20* are the 20 features chosen by PCA.

Bajaj et al.[54] used emperical mode decomposition (EMD)[14] features as inputs for a Least Squares Support Vector Machine (LS-SVM) driven seizure classifier. The data was sourced from 100 23.6s channel epochs drawn from 5 subjects. EMD separated the nonlinear and non-stationary components of the EEGs into intrinsic mode functions (IMFs). The two dominant IMFs, amplitude modulation and frequency modulation, produced a peak sensitivity and specificity of 100% while averaging 94% sensitivity

[13] area, normalized decay, frequency band power, line length, mean energy, average peak/valley amplitude, normalized peak number, peak variation, root mean square, wavelet energy, and zero crossings

[14] A detailed review of EMD is omitted, but if interested the work of Huang et al.[154] introduced technique and its applications.

and specificity over the dataset by using a common supervised ML algorithm in SVMs. As the classifier was not knew, the success of this work was likely driven by the use of EMD features or qualities of the 5 subject dataset. This was a recurrent problem with EEG algorithm development of unique datasets and unique features obscuring the cause of classification improvement.

The alternative to diverse features and datasets was to test range of algorithms. This is what Acharya et al.[152] did by focusing on six supervised ML algorithms: Fuzzy Sugeno Classifier (FSC), SVM, KNN, Probabilistic Neural Network (PNN), DT, and Naive Bayes Classifier (NBC), and one unsupervised ML algorithm: GMM. This was a better approach than Bajaj et al.'s as it provides multiple reference points on a constrained dataset. These six algorithms used four different types of entropy calculations as features: Approximate Entropy (ApEn)[155], Sample Entropy (SampEn)[156], and S1 entropy and S2 entropy[157]. Distinct epochs were drawn from 5 healthy and 5 epilepsy subjects that produced 200 healthy, 200 preictal, and 100 ictal artifact free single channel 23.6 second epochs.

Table 2.17: Classification accuracy of entropy based feature sets for various classifiers.

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)
FSC	98.1	99.4	100
SVM	95.9	97.2	100
KNN	93.0	97.8	97.8
PNN	93.0	97.8	97.8
DT	88.5	98.3	91.1
GMM	95.9	98.3	95.6
NBC	88.1	94.4	97.8

The sensitivity and specificity the algorithms were similar, table 2.17, but the best accuracy was achieved by the FSC classifier. The separability of the trained seizure states, table 2.18, produced a p-value less than 0.0001 for each entropy. However, it was hard to assess the strength of the individual algorithms given the small size of

the dataset and the natural discrimination strength of the features. The strong performance across all the algorithms suggested the results were driven by the features, but the dataset was, again, too small to have known for certain.

Table 2.18: The level of entropy for each feature with respect to the classified seizure state.

Class	Normal	Pre-ictal	Epileptic
ApEn	$2.2734 \pm 3.320 \times 10^{-2}$	1.8650 ± 0.331	1.9325 ± 0.215
SampEn	1.3130 ± 0.120	0.99332 ± 0.189	0.92628 ± 0.139
S1	$0.57012 \pm 7.120 \times 10^{-2}$	$0.47208 \pm 6.149 \times 10^{-2}$	0.48325 ± 1.55
S2	$0.76827 \pm 3.125 \times 10^{-2}$	$0.68072 \pm 3.790 \times 10^{-2}$	$0.73184 \pm 4.555 \times 10^{-2}$

Ghosh-Dastidar et al. [25] benchmarked a novel wavelet-chaos-neural network Levenberg-Marquardt Backpropagation Neural Network (LMBPNN) classifier against the same data as Acharya et al. They picked features (standard deviation, correlation dimension, and largest Lyapunov exponent) that were specific to each frequency band and grouped them together into various band specific sets. The epochs were evaluated by supervised techniques (Radial Basis Functional Neural Network (RBFNN) and LMBPNN), an unsupervised technique (k-means clustering), and statistical discriminant techniques (Quadratic Discriminant Analysis (QDA) and LDA using Euclidean and Mahalanobis distance metrics).

The various combination of band-specific features sets were used to resolve an optimal set for the classifiers. These tests provided an exhaustive analysis of the relationship between algorithm and feature set performance, which was frequently lacking in other research. However, the best performance resulted when using a mixed-band feature set. The impact of feature set on the performance of LMBPNN is seen in figure 2.7.

The classification performances were similar so only the maximum accuracy was reported in table 2.19 and even these cover a wide range. Overall, the proposed

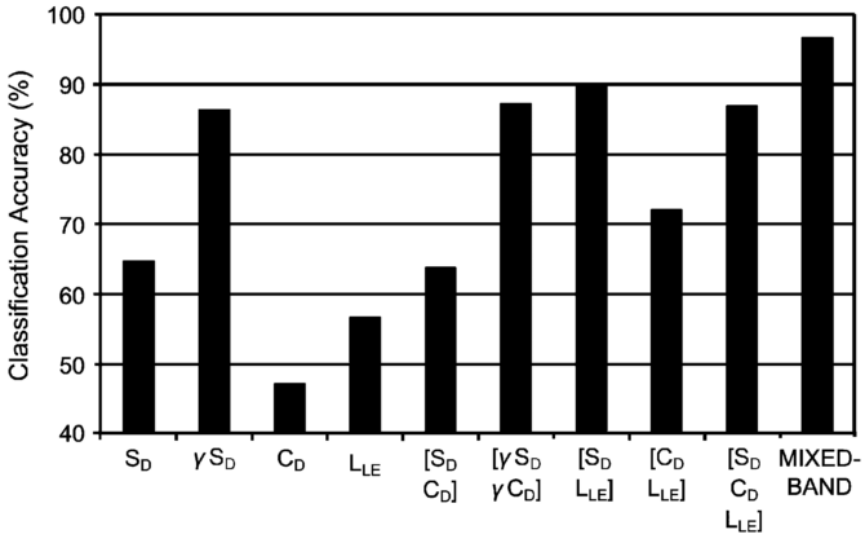


Figure 2.7: When iterating over the available feature sets, the performance of LMBPNN responds differently for each combination. The Greek letters indicate the band specific features (S_D is standard deviation, C_D is correlation dimension, and L_{LE} is Lyapunov exponent). The mixed-band feature set uses the band independent S_D and C_D with $\alpha S_D C_D L_{LE}$, $\beta S_D C_D$ and $\gamma S_D C_D$.

LMBPNN provided the strongest peak performance, but it relied on a large mixed-band feature set. This suggested that the features were the driving force of classification, and yet only some of the algorithms were able to adequately used them. Again, the difficulty in improving performance appeared to stem from being able to model the phenomena in a meaningful way for the chosen classifiers.

2.2.2.3 Unsupervised Algorithms

Unsupervised ML algorithms differed from supervised ML algorithms in that they do not require labeled data. This has made them a historically useful as starting point when knowledge of a domain was limited. Their decision surfaces were created directly through the data which removed any bias present in the labeling, but traded it for bias in the datasts. This also meant unsupervised approaches worked best when

Table 2.19: The table reports the maximum accuracy achieved by each algorithm given on a single or (*) mixed-band feature set.

Algorithm	Maximum Accuracy (%)
k-means	59.3
LDA w/ Euclidean	79.6
LDA w/ Mahalanobis	84.8
QDA	85.5
RBFNN	76.5
LMBPNN	89.9
QDA*	93.8
LMBPNN*	96.7

operating on datasets with enough data to represent each class of interest. Thus without an equitable distribution of data, classes may be ignored or poorly modeled leading to weak classification performance.

Given the need for large and diverse datasets and improvement to supervised classification techniques, the use of unsupervised classification of EEG recordings has diminished. However, unsupervised algorithms endured given their ease of use and ability to produce benchmarks for their supervised counterparts. Acharya et al.[152] showed GMM produced competitive accuracy and sensitivity, but not specificity to their tested supervised methods in table 2.17. Alternatively, there were cases where it peformed much worse such as Ghosh-Dastidar et al.'s [25] k-means clustering of table 2.19. As unsupervised algorithms relied on the dataset more than supervised techniques, such contrasting performances were common in the EEG literature.

Gabor et al.[158] tested a single unsupervised algorithm, a self organizing map (SOM)[15] NN, for seizure detection on 24 recordings from 22 subjects. The algorithm was trained to classify seizures from features produced by a wavelet transform using

[15]A detailed review of SOMs was omitted, but if interested the work of Kohonen[159] formalized the implementation. This technique attempts to mimic the structure of the brain by parsing the data in an unsupervised fashion to create a flat, two dimensional, map linking elements of the data together.

4s epochs built from the 10 channels of each recording. A separate feature set using 8s epochs was used, but the duration was found to be too long as it masked out shorter seizures.

In total 62 seizures were captured from the 24 recordings of which the algorithm detected 56 (90%). The average false positives per hour (0.71) produced more false positives than true positives given the average recording duration of 22.02 hours. As discussed previously, unsupervised techniques are sensitive to the distribution of the training data which manifested in this case as poor false alarm rates. In addition, the age range (<1 to 43 years old), small training set (5 of the 24 recordings), and epoch duration were all factors working against the algorithm.

Not all unsupervised algorithms focus on classifying the data, as some are deployed for dimensionality reduction. One such approach was the use of of unsupervised LDA in areas where clinicians' skills were weaker such as BCI [83]. LDA is within the realm of factor analysis (FA), as are ICA and PCA. A detailed review of FA techniques is given in Section 2.4.1 and what follows now touches on their use in EEG applications.

Vidaurre et al. [83] used three flavors of LDA to enhance BCI performance over four datasets[16]. The experiments focused on developing an unsupervised solution to transitioning between training and feedback sessions of BCI tasks. Each version of LDA focused on a different aspect of the features: LDA-I, targeted changes in the pooled mean (PMean) between the features of the training and feedback data, LDA-II incorporated updates to the covariance matrix with PMean, and LDA-III scaled the mean and covariance using CSPs. These techniques were compared against a supervised version of LDA to determine the strength of the unsupervised techniques.

[16]The first dataset was comprised of 19 sessions recorded from 10 subjects performing motor imagery tasks. The second dataset consisted of 80 subjects performing 75 motor imagery trials with calibration. The third dataset involved 7 quadriplegics attempting to move use a BCI mouse. The final dataset was a repeat of the second dataset without any calibration for the users.

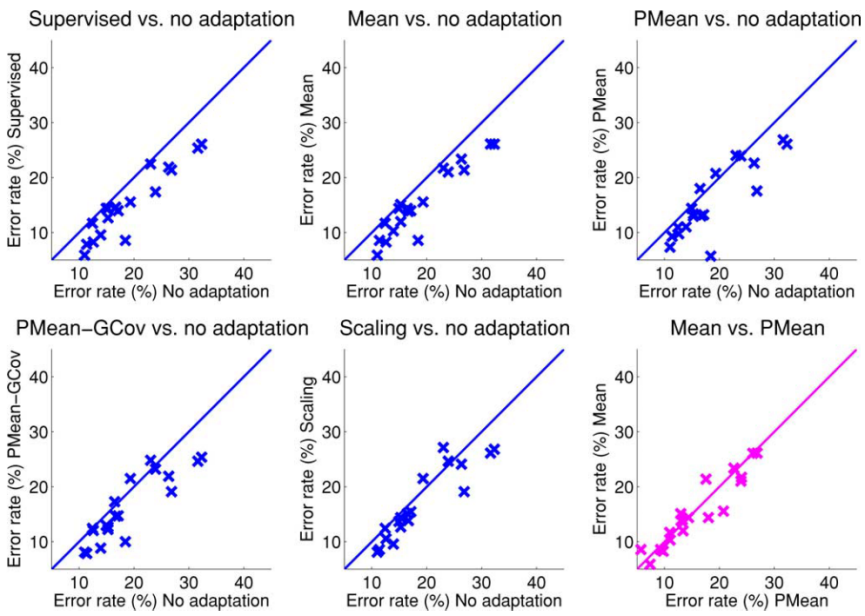


Figure 2.8: The comparative error rates between the supervised and unsupervised adaptation techniques through changes in the error rate. The pink plot shows the difference between a labeled, mean, and unlabeled, PMean, classification.

Unsupervised techniques main strength resides in their simplicity when compared to ever advancing supervised techniques. Their evaluation was frequently used to evaluate the performance gain versus increased complexity and requirements. This made head to head comparisons, figure 2.8, critical to development of both type of algorithm. Using the first dataset, the supervised algorithms slightly out performed the unsupervised algorithms. On the second, larger, dataset the PMean based algorithms met or exceeded the performance of the state of the art supervised approaches during feedback. The unsupervised technique exhibited robustness as a class was removed from BCI feedback and outperformed the supervised algorithm in figure 2.9. These results were important because clinicians seldom label BCI datasets and it showed the trade off between supervised and unsupervised may not be that advantageous. This

was especially true in this instance as BCI recording sessions were rarely annotated by clinicians, but often as dynamic as seizure or sleep sessions.

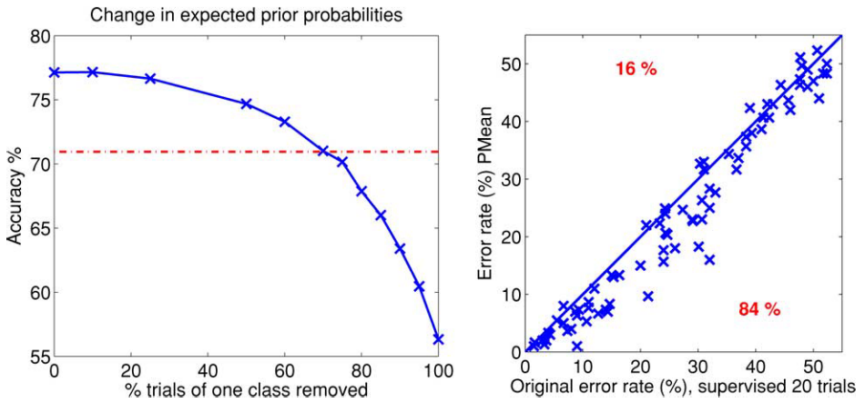


Figure 2.9: Performance on feedback data after training for supervised adaptation and unsupervised PMean adaptation. The (left) impact of removing one class from the feedback dataset for the supervised algorithm (red line) and unsupervised algorithm (blue line). The (right) error rate between the two algorithms during the online feedback experiments.

2.2.3 Bio-metric Applications

The use of EEG recordings as a means of bio-metric identification was not dominant area of research, but has begun to rapidly advance [105]. Initial efforts focused on being able to discriminate EEG behavior between individuals and between different brain conditions [160]. This work did not have discrete waveforms to find or frequency ratios to calculate, but instead relied on direct comparison between subjects. Stassen [161] developed computerized methods, borrowed from speech recognition, to recognize normal and schizophrenic individuals based on their EEG spectral pattern. The style of this approach, finding dominant properties in subject epochs, remained in use [59] and was the best corollary to the research proposed in this dissertation.

Advancement of EEGs as a bio-metric tool focused on the statistical properties of each subject [105]. This detached it from the dominant research trends that were

reliant on clinical annotations [103]. This made bio-metric applications open-ended as they do not, and in most cases cannot, rely on previously developed feature sets or decision surfaces built from clinical annotations. Researchers were therefore on their own to find features and testing protocols leading to a variety of approaches not seen elsewhere [63, 65, 104, 162, 163].

The initial efforts by VanDis et al.[160] and Stassen [161] focused on subjects at rest with eyes closed and open. Similar experiments contiuned to be carried out [33, 64, 65] but with their aims updated to optimize the accuracy and speed of subject verification as a function of features and channels. Active state recordings, when subjects performed mental tasks, such as imagined hand movements [68, 103], imagined speaking syllables [66], or reading text[39] underwent similar channel and feature optimization testing.

Active and resting based data analysis suggested that the qualities of subject authentication and identification existed regardless of brain state. Other works went as far as suggesting a genetic basis underlies this separability [37, 67]. While interesting, the genetics of brain uniqueness expanded beyond the scope of this work. However, by focusing on the techniques and results of active and resting based data studies comparisons could likely be drawn between the structured waveform based annotations of artifacts, seizures, and sleep.

2.2.3.1 Resting Recordings

The work of La Rocca et al.[33, 64, 65] focused on developing a novel set spatial and temporal patterns as features to improve subject recognition. Brigham et al.[66] used data with imagined activities to test applications of subject identification during mental tasks. These studies represent the adaptation of techniques the worked for

other EEG classification tasks, spatial and temporal patterns for BCI and mental tasks for attention/focus/workload performance and ERPs.

In [33] electrode sets of 2, 3, and 5 from the 56 recorded channels were used to find a lower-bound on the number of required channels. The approach used autoregressive stoichastic modeling and polynomial regression to match 3 second epochs broken into features through the 6 standard EEG bands. Classification performance varied as a function of electrode set and the EEG band used. Increasing electrodes trended with an improvement in classification performance. However, regardless of the number of electrodes the alpha band provided the strongest classification accuracy. Performance peaked at 98% classification accuracy when using the alpha, beta, delta, and gamma bands for 5 channel sets. The best single band performance (83%) was seen using only the alpha band across 5 channels.

They followed up this work with 'bump' modeling to reduce the amount of data from the 10-20 layout into a parametric model [65]. These bumps were filters that enabling sparse encoding. This generated vectors to control the mapping/weights of the bumps scale the features of the data. These vectors were then classified with LDA based upon features generated from groups of three channels drawn from the six standard EEG bands. The training and testing sets were curated to provide overlapping frames, *jointed*, and without overlapping frames, *disjointed*. This distinction highlighted the impact of frame overlapping with the beta band producing a classification accuracy of 95% for jointed and 74% for disjointed. The alpha band resulted in similar classification accuracy with 96% for jointed and 67% for disjointed. In all bands the jointed feature sets outperformed their disjointed counterparts.

Their final work focused on spatial patterns generated from 1s PSDs epochs from different regions of the brain [64]. This deviated from their earlier attempts at reducing the amount of data through feature and individual channel reduction. Instead

it grouped channels together to develop a statistical approach to subject verification using the PhysioNet Database dataset. Classification was carried out by building Gaussian mixtures based upon the distributions of the PSDs. These mixtures were evaluated via a Mahalanobis Distance (MD) classifier to determine likelihood of similarity between subjects. Using the results for each region of the brain, classification accuracy reached 100% for identifying subjects.

2.2.3.2 Active Recordings

In Marcel et al.[103] a nine subject dataset was classified based upon their brain activity during three mental tasks. These tasks required the subjects to imagine carrying out prescribed actions: moving their left hand, moving their right hand, and speaking words with a common leading letter. The feature set was built from 0.5s 50% overlapping epochs of PSDs. These PSDs were spatial filtered over the 10-20 electrode configuration with a surface Laplacian function. The features were given to a GMMs which produced baseline models for subject verification. Evaluation scores were reported as half total error rate (HTER) generated from the false acceptance rate (FAR) and false rejection rate (FRR).

$$HTER = \frac{FAR + FRR}{2} \tag{2.2-1}$$

The results, table 2.20, of the left and right hand authentication of the subjects suggested performance was dependent on the number of Gaussian mixtures used in the modeling process. This experiment used a large datasets which was collected from the subjects over a three day period. Results using smaller subsets of the dataset showed the imaging word task authentication lagged compared to that of the hand tasks.

Table 2.20: The FAR, FRR, and HTER of imagined hand tasks as a function of Gaussian mixtures for Marcel et al.'s experiments.

Mental Task	Num. Gaussians	FAR	FRR	HTER
Left	4	18.6	32.3	25.4
	8	23.8	25.15	24.5
	16	19.3	19.65	19.5
	32	13.7	24.9	**19.3**
Right	4	18.4	40.5	29.4
	8	20.6	29.5	25.0
	16	15.0	23.6	**19.3**
	32	13.0	30.15	21.6

In Fraschini et al.[68] phase synchronization was tested as a feature set for identifying subjects. The dataset used the 109 PhysioNet Database subjects' resting eyes closed and resting eyes trials. The features were generated from the standard EEG frequency bands and segmented into 12s non-overlapping epochs. Finding the phase lag index (PLI) relationship between all the channels of an epoch produces distinct mappings between subjects. These topologies were reduced via Eyes Closed (EC) to produce a feature vector for each epoch. The Euclidean Distance (ED) between each feature vector was the decision surface used to assert the similarity between the subjects for each frequency band.

Table 2.21: EER of phase synchronization based subject verification

Band	REO EER (%)	REC EER (%)
Gamma	4.4	6.5
Beta	10.2	16.9

Brigham et al.[66] tested a similar subject identification protocol using two unique datasets. One source of data came from Visually Evoked Potentials (VEPs) in 120 alcoholic and non-alcoholic subjects. The other ws sourced from 6 subjects uttering two syllables, /ba/ and /ku/. Artifacts were removed from each set and processed into PSDs of their respective trial lengths, 1s for the VEP and 10 seconds for the syllables.

Using SVMs and KNNs the classification accuracy of each algorithm was averaged from 4-fold cross-validation. After artifact removal the VEP data set contained 9,596 trials for the 120 subjects and 3,787 trials for the 6 syllable subjects.

On the VEP dataset the SVM achieved 98% accuracy and KNN achieved 93% accuracy, both had a 95% confidence interval. The syllable dataset achieved higher accuracy, 99% with SVM and 98% with NN with both, again, at a 95% confidence interval. The strong performance across both datasets indicated the techniques and feature sets worked well on a fundamental level. However, the diminutive number of syllable subjects was not compelling and should have likely be run with more subjects.

In Gui et al.[39] a more contemporary ML technique, Artifical Neural Network (ANN) using feed-forward, back-propagation, and multiplayer perceptron, was used to identify subjects. Their dataset consists of the 6 mid-line channels {Fpz, Cz, Pz, O1, O2, and Oz} of 32 subjects undergoing VEPs. The channels were bandpass filtered, 0Hz to 60Hz, before wavelet packet decomposition (WPD) produces the final three features of mean, variance, and entropy for each 1.1 second epoch. Four experiments are carried out, but only two were of interest in subject classification: (S1) finding a single subject from the set of 32 and (S2) matching all 32 subjects against each other simultaneously. The other experiments consisted of a one versus all classification (S3) and separating small groups of subjects from each other (S4). For S1 the highest accuracy of 10% occurred with 5 neurons and the worst accuracy of 5% occurred with 10 neurons. S2 produced better results with a highest accuracy of 94% with 45 neurons and a worst accuracy of 70% with 30 neurons.

Here the fundamental issues of ML were exposed in terms of dataset source, pre-processing, feature selection, algorithm, and classification task. Across the range of experiments presented nearly every single one used different data or different features or different algorithms. The result was a lack of comparison points from which to

drawn definitive conclusions about EEGs data and features or their classification. While many of these experiments produced acceptable results, little was gained and many were often confirming ideas already well documented instead of expanding the knowledge base.

2.3 Identity Vectors

At the center of this dissertation was the introduction of I-Vectors to the EEG classification community. I-Vectors are mathematical models that were designed to reduce the dimensionality of UBMs [117]. UBMs served to reduce a dataset of f-dimensional feature samples into C mixtures of f-dimensional GMMs. Following this, I-Vectors can then be created by enrolling distinct samples into a modeling process involving the UBM and a TVM. The TVM is generated from the enrollment samples and served to constrain the contributions of each mixture within the UBM. Finally, those I-Vectors were evaluated against testing I-Vectors, built from testing samples and the same TVM used to produce the enrollment I-Vectors. This evaluation resolve the distance in the l-dimensional distance between them.

As the technique is entirely data dependent, it could be altered to measure similarities between epochs, channels, individuals, or groups of individuals. I-Vectors were developed originally as an extension of a speech processing method called joint factor analysis (JFA) which split utterances into separate models for speaker, channel, and context [164]. In contrast, I-Vectors collapse those three models into just one.

The principal I-Vector equation is

$$M \approx m + Tw \tag{2.3-2}$$

where M is the feature space of the data, m is the UBM, T is the TVM and w is the I-Vector itself. The specific data used to build the UBM m is referred to as the training data. Once m and T have been defined, they can be used in concert with alternate enrollment targets of size S and testing data sets M to create data-specific I-Vectors, w.

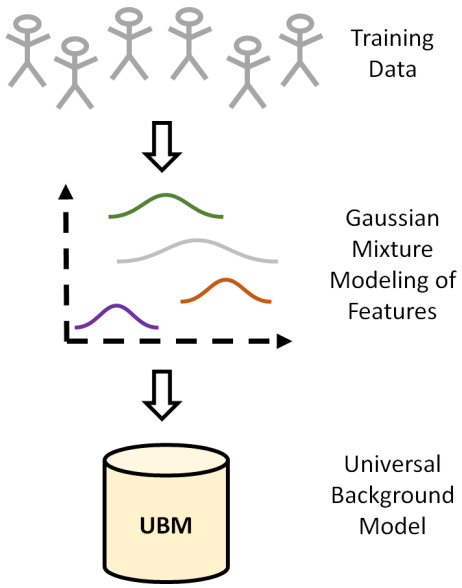


Figure 2.10: Training data was used to construct C independent Gaussian mixtures over the f dimensional feature space. This transformed the training data into C mixtures each with f means and variances. Taken as a whole these c mixtures were the UBM in addition to a mixture weight parameter. Ultimately this c mixture UBM served as the basis for developing a TVM and the associated I-Vectors.

A typical I-Vector use case might involve determining whether an EEG from a new patient should be diagnosed as epilepsy. First, a large randomized collection of training data drawn from a diverse set of subjects would be used to build a UBM, figure 2.10. Then, sub-populations of data from known healthy and epileptic patients would be used to construct an enrollment dataset. This enrollment dataset would be used to resolve a TVM and produced enrollment I-Vectors related to the enrollment subjects.

Finally, the new patient's data would be used with the TVM to construct their I-Vectors. Then evaluations between the enrollment and target I-Vectors would inform which population they were more likely to match with, figure 2.11. Depending on the choice of enrollment and test data, I-Vectors can automatically search for across channels, times, medical conditions, medications, and even entire subjects.

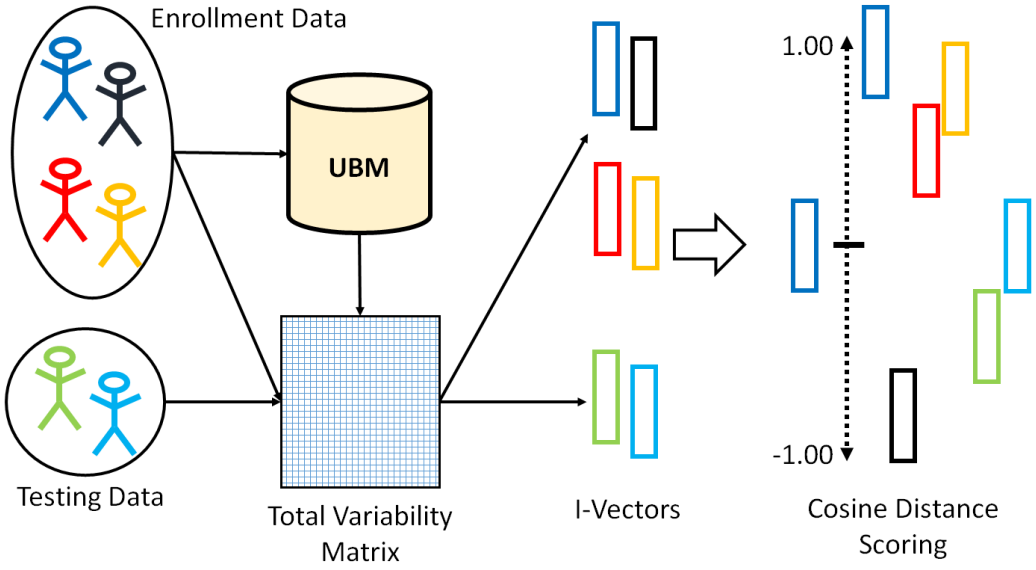


Figure 2.11: Using the UBMs as an initialization, the enrollment and training data are transformed into I-Vectors. This process is reliant on the creation of the total variability matrix randomly generated from the variances of the UBMs and refined by adaptation towards the means of the UBMs. The resultant I-Vectors are pairwise evaluated to find the CD between them to rank their similarity.

A UBM models the f-dimensional features by representing them with C independent Gaussian mixtures [165]. In general, increasing the number of mixtures captures more nuance, thereby potentially strengthening discrimination. The UBMs provide dimensionality reduction by taking a training dataset of L epochs each with f features each down to C mixtures of f features. As each feature has a mean m, variance σ, and weight ρ, reduction benefits are seen when $L > 3C$. The UBMs can be characterized

according to:

$$\Omega_{c=1...C} = \begin{cases} m(c) \\ \sigma(c) \\ \rho(c) \end{cases} \tag{2.3-3}$$

Each parameter is a vector of length f representing a given feature. Each I-Vector is the result of the expectation maximiation (EM) of the available UBM and M.

The I-Vectors are of length $l = Cf$ with many residual elements and are frequently further reduced by the speech community via LDA. This process requires that the final I-Vector length be one less than the number of subjects S or less given the constriants imposed by LDA's algorithm. Thus I-Vectors final length, $l = min(S - 1, Z)$, is frequently controlled by S or Z, where Z is on the order of 100s. The final I-Vectors therefore represent a very dense and robust abstraction to an l dimensional space. Within this space the similarity between two I-Vectors can be found via any metric evaluation, often CD.

2.3.1 Mathematics

The major components and steps to producing I-Vectors are outlined in reverse starting with the resultant I-Vectors and ending with the original JFA technique. This includes sections on the TVMs, UBMs, maximum a priori (MAP), and GMMs.

2.3.1.1 I-Vectors

The critical component of equation 2.3-2, is the TVM T. An evolution from the eigenvoice matrix used in JFA, it captures all of the variances present in the UBMs. Generating T from training data requires an iterative EM approach reliant on feed-

back from the produced I-Vector w.

$$T = \begin{bmatrix} T_1 \\ \vdots \\ T_C \end{bmatrix} = \begin{bmatrix} A_1^{-1} * K_1 \\ \vdots \\ A_C^{-1} * K_C \end{bmatrix} \tag{2.3-4}$$

The matrices of A and K represent the updated mean and variance of T. These updates are driven by w and T along with the static values of $N, \hat{F}$, and Σ. The superscript H represents the Hermitian transpose.

$$A_c = \sum_{s=1}^{S} N_s(t) w^{-1}(t) \tag{2.3-5}$$

$$K_c = \sum_{s=1}^{S} \hat{F}_c(s) * \left(w^{-1}(s) * T^H * \Sigma^{-1} * \hat{F}_c(s) \right)^H \tag{2.3-6}$$

The estimation of w uses T a $Cf \times Cf$ matrix. This matrix is formed from the Baum-Welch (BW) statistics $\hat{N}$ and $\hat{F}$, an $l \times l$ identity matrix I, and a model of the UBM variances Σ. As the models are all independent Σ is a diagonal $Cf \times Cf$ matrix of the true variances from the UBMs where as the BW statistics are estimations of the mean N and variance F.

$$w(s) = \left(I + T^t \Sigma^{-1} \hat{N}(s) T \right)^{-1} T^t \Sigma^{-1} \hat{F}(s) \tag{2.3-7}$$

The BW 0^{th} (N) and 1^{st} (F) order statistics are generated from the evaluation of the UBMs against the L epochs in the training data. The higher order statistic must be offset by the preceding orders resulting in a centered 1^{st} order statistic $\hat{F}$. Each statistic models the f features in each of the C mixtures resulting in $C \times f$ matrices. Each epoch, e, from the full epoch set $t = 1...L$ is evaluated to generate

initial probabilities based on Ω for N and F.

$$\hat{N}(s) = \begin{bmatrix} N_1(s) & & \\ & \ddots & \\ & & N_C(s) \end{bmatrix} \tag{2.3-8}$$

$$\hat{F}(s) = \begin{bmatrix} \widetilde{F}_1(s) \\ \vdots \\ \widetilde{F}_C(s) \end{bmatrix} \tag{2.3-9}$$

$$\widetilde{F}_c(s) = F_c(s) - N_c(s)m_c \tag{2.3-10}$$

$$N_c(s) = \sum_{t=1}^{L} P(c \mid e_t, \Omega) \tag{2.3-11}$$

$$F_c(s) = \sum_{t=1}^{L} P(c \mid e_t, \Omega)e_t \tag{2.3-12}$$

This process resolves a suitable T after approximately twenty iterations of equations 2.3-4 to 2.3-7. Notice that equations 2.3-8 to 2.3-12 are needed only once to generate T. Creating I-Vectors from the enrollment and testing data follows equation 2.3-2 in a modified form. The resultant I-Vector w will be a l row vector where l is a length defined during the creation of the initial estimate of T.

$$w = (M - m)T^{-1} \tag{2.3-13}$$

The number of I-Vectors produced is based upon the enrollment targets h and testing queries q, producing data on the order of $(h + q) \times l$. Therefore dimensionality reduction will not be significant if the data is partitioned such that $h + q \equiv L$.

The I-Vectors are finalized after applying LDA to control for dependencies in the data. This process reduces their length from l to $l = min(S-1, l)$ elements based upon the transformation matrix produced by the LDA. There are other approaches to normalize the I-Vectors aside from LDA which can be reviewed elsewhere [166]. These final I-Vectors can be compared pairwise using CD to determine similarity between enrollment targets and testing queries.

$$cos(\Theta_{w_1,w_2}) = \frac{w_1^t w_2}{\|w_1\| * \|w_2\|} \tag{2.3-14}$$

2.3.1.2 Total Variability Matrix

After the development of JFA it was discovered that the iterative modeling process was not perfect at separating speaker, channel, and residual effects[166]. In fact the eigenchannel space was collecting information related to the subject when operating on specific utterances. JFA was still considered state of the art, but its performance could be challenged by the total variability space. This space, formally the TVM, was produced by using the first iteration of JFA to generate a low-dimensional speaker- and channel-dependent matrix. As this matrix is the key component in generating I-Vectors a detailed decomposition of its construct and applications is necessary.

The initial form of the T is $f \times C$, GMMs by features, shown in equation 2.3-15. These parameters were dependent on each other and the training data. The speech community uses a definitive feature set [137], Mel Frequency Cepstral Coefficients (MFCCs), which evolved over time to become the gold standard [167]. This makes determining the number of features straightforward. Settling on an acceptable number of mixtures for the GMM was more difficult given the trade-offs between classification and computational performance[168, 169].

In many studies the number of mixtures is on the order of a base 2 number, often being set to at least 2048 mixtures[170, 171]. The optimization for the number of mixtures was dependent on the best performance, but limited by the dimensions of the training data. Given a number of subjects S each providing u utterances the number of mixtures C would need to be less than $S * u$ to prevent over-fitting.

$$\begin{bmatrix} M_1 \\ \vdots \\ M_f \end{bmatrix} = \begin{bmatrix} m_1 \\ \vdots \\ m_f \end{bmatrix} + \begin{bmatrix} T_{1,1} & \dots & T_{1,C} \\ \vdots & \ddots & \vdots \\ T_{f,1} & \dots & T_{f,C} \end{bmatrix} * \begin{bmatrix} w_1 \\ \vdots \\ w_C \end{bmatrix} \tag{2.3-15}$$

Critically, the TVM was not implemented to mimic utterances, but to map them instead. The technique allowed I-Vectors to be the weights controlling the inclusion of a column of features. In this manner it was possible that one column may contain the dominant features of a low pitched voice and a high pitched voice. If each of the C columns of T represent a unique component of the speakers, then the I-Vector w would be binary. More likely is that the characteristics are spread across mixtures since emergent properties of speech are parameterized via the MFCCs.

Advancing this approach to EEGs may produce a reasonable algorithm for discrimination, but also allow for an understanding of why the discrimination occurs. This is entirely dependent on the chosen features, which are well established for speech, but still open for EEGs. Using a non-linear variation of MFCC maintains the parameterization providing a closed set of features. With features bounded, experiments can then focus on finding an optimal size GMM for the UBM of EEGs.

Working down this chain, further incremental improvements can be made while gaining insight into the discrimination and grouping of EEGs in an unsupervised algorithm. While speech already knows the principal modes of their data, how to

separate consonants, vowels, words, genders, and ages, such techniques do not meet the needs of the EEG community.

2.3.1.3 Universal Background Models

As mentioned previously UBMs are sets of GMMs created from the features of continuous signals. The GMMs contextualize the varied speech signal segments as independent feature distributions regardless of the spoken text [165]. This technique is suited to the problem of speaker recognition where the goal is to match subjects irrespective of data content. As this process is reliant on the likelihoods of features for a given model or subject sample, it can be used in an unsupervised manner to match and/or separate subjects.

The GMM represents the core component of the UBMs which in turn makes them critical to the performance of I-Vectors. Sets of Gaussian distributions (M) can be represented with a mean (μ) and co-variance (Σ) drawn from each measurement or feature of the D-dimensional raw continuous data [119]. This allows a likelihood calculation equation given a D-dimensional sample x to compare against the model,

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} w_i g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2.3-16}$$

where x,μ, and Σ are vectors of length D and w_i corresponds to the weight of each mixture component where $\sum_{i=1}^{M} w_i = 1$. The calculated likelihood provides an unsupervised estimation of the sample relating to the given model(s).

The λ component of $p(\boldsymbol{x}|\lambda)$ represents the GMM and associated parameters: w_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$. While the previous equation does not assign a subscript to λ there would be U GMMs which comprise the fully formed UBM. Just as each GMM attempts to

determine the underlying states of the data, the UBM requires depth to account for each class of signal.

As an example suppose one wants to know if the weather on a given day will require a heavy coat, a light coat, a raincoat, or no coat. If the temperature is below 45°F a heavy coat is desired and if the temperature is above 70°F no coat is necessary. In between these two temperatures a light coat may be necessary, but only if the day will be windy. At the same time, at any temperature above 45°F with high humidity levels should warrant wearing a raincoat.

The GMM representing raincoat would have a large variance for wind and temperature, but a small variance for humidity. The temperature means of heavy coat, light coat, and no coat would be unique. However, light coat and no coat would have a similar mean and variance for humidity and overlapping distributions for wind. Meanwhile, the heavy coat model would be insensitive to anything aside from temperature.

The weather conditions (humidity, temperature, and wind) become the three features modeled by the GMMs. Once four, or more, models are created they each categorize the required jacket. This full set becomes the UBM that provides a basis for evaluation of each day's weather. Given a weather report, the UBM would provide the likelihood of each jacket being the correct answer.

To calculate the likelihood for a multivariate normal distribution the follow equation is used, represented as the function $g(\boldsymbol{x}|\mu_i, \boldsymbol{\Sigma}_i)$ from the prior equation,

$$g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{\exp\Big\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\Big\}}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \tag{2.3-17}$$

From these equations estimations of underlying modes of the data can be found from which to build a suitable model. Two important assumptions are made in this

process, the first is that each Gaussian mixture is independent of the other mixtures and the second is that the underlying modes can me adequately modeled with normal Gaussian distributions. These mixtures are therefore representing a unique hidden set of generators/states that create the resultant signal. Given that the number of hidden states is unknown, GMMs may produce mixtures with marginal weights or mixtures with redundant attributes.

2.3.1.4 Maximum A Posteriori Parameters

With a UBM in place it is possible to tune the model toward specific subjects. The estimation of a subject specific model from a UBM is called MAP estimation[119]. Just as with a UBM, the statistics (weight, mean, and variance) of the subject are found from their data $S = \boldsymbol{s}_t, ..., \boldsymbol{s}_T$. These expectations are derived from the prior model found from the UBM, but operating on the subject specific data.

$$n_i = \sum_{t=1}^{T} \Pr(i|\boldsymbol{s}_t, \lambda_{\text{prior}}) \tag{2.3-18}$$

$$E_i(\boldsymbol{s}) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|\boldsymbol{s}_t, \lambda_{\text{prior}}) \boldsymbol{s}_t \tag{2.3-19}$$

$$E_i(\boldsymbol{s}^2) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|\boldsymbol{s}_t, \lambda_{\text{prior}}) \boldsymbol{s}_t^2 \tag{2.3-20}$$

These are then able to adapt each i mixture's weight, mean and variance. The amount of adaptation is based on the expectations and a chosen relevance factor r^ρ.

$$\hat{w}_i = \left[\frac{\alpha_i^w n_i}{T} + (1 - \alpha_i^w) w_i\right] \gamma \tag{2.3-21}$$

$$\hat{\boldsymbol{\mu}}_i = \alpha_i^m E_i(\boldsymbol{s} + (1 - \alpha_i^m) \boldsymbol{\mu}_i \tag{2.3-22}$$

$$\hat{\boldsymbol{\sigma}}_i^2 = \alpha_i^v E_i(\boldsymbol{x}^2) + (1 - \alpha_i^v)(\boldsymbol{\sigma}_i^2 + \mu_i^2) - \hat{\boldsymbol{\mu}}_i^2 \tag{2.3-23}$$

The adaptation coefficient is most often constant for all three statistics, but given unique labeling allowing for decoupling if necessary.

$$\alpha_i^{w,m,v} = \frac{n_i}{n_i + r^\rho} \tag{2.3-24}$$

These new statistics not only provide subject specific models, but present a new set of models for discrimination. An example of this process is shown in figure 2.12. The models themselves can be compared against each other to determine similarity in addition to evaluating them against new data samples.

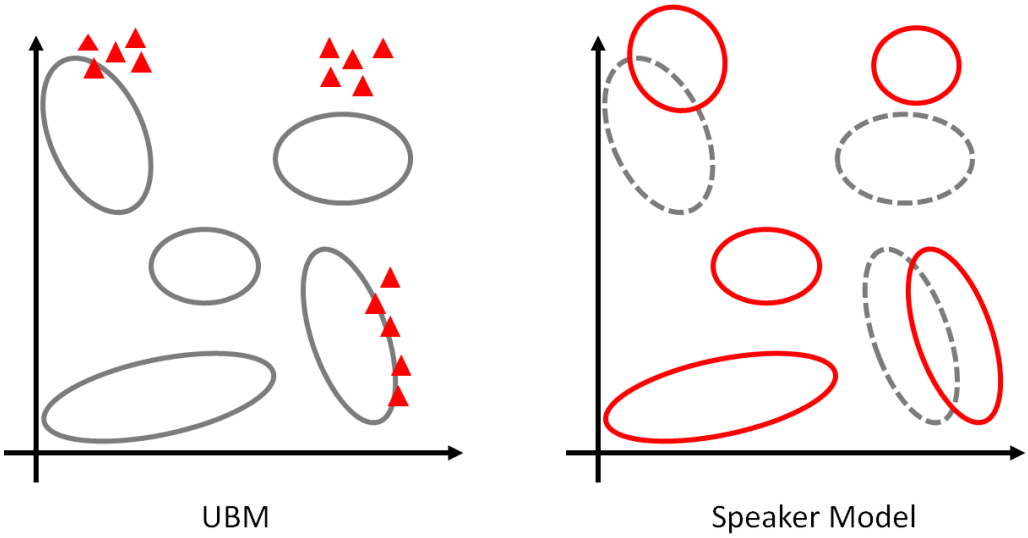


Figure 2.12: Results of MAP estimation when speaker data, red triangles, is applied to a UBM, gray mixtures.

2.3.1.5 Gaussian Mixture Models

Understanding how GMMs produce likelihoods for a given data sample $\boldsymbol{x}$ informs how each mixture's λ is produced. The more accurate the parameters of λ are for a given

GMM, the more insightful the resultant likelihoods. However, unless the parameters are known outright they must be deduced empirically. One of the more prevalent techniques for parameter estimation is maximum likelihood estimation (MLE)[172].

The MLE attempts to find a distribution that maximizes each of the T training vectors $X = \{x_1, ..., x_T\}$

$$p(X|\lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\lambda) \tag{2.3-25}$$

this equation assumes that each component of the distribution is independent. This often turns out to be untrue, but is a necessary assumption to provide a functional solution. This function is non-linear as the product of all the training vector evaluations allows for one worsening likelihood to diminish any improvements gained from the remaining vectors. To avoid this problem, a variant of EM can be used to estimate the parameters for each feature independently. This helps isolate the features, in the event that they are not independent, and provides the ability to directly improve the overall likelihood on a feature by feature basis.

With this each parameter of λ can be estimated in an iterative manner with the following equations

$$\bar{w}_i = \frac{1}{T} \sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda) \tag{2.3-26}$$

$$\bar{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t}{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)} \tag{2.3-27}$$

$$\bar{\boldsymbol{\sigma}}_i^2 = \frac{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t^2}{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)} - \bar{\mu}_i^2 \tag{2.3-28}$$

these three equations provide updated values for the weights, means, and variances that can feed the next iteration of the EM algorithm. The *a posteriori* probability

Pr is found with the following equation

$$\Pr(i|\boldsymbol{w}_t, \lambda) = \frac{w_i g(\boldsymbol{x}_t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{k=1}^{M} w_k g(\boldsymbol{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \tag{2.3-29}$$

2.3.2 Success in Speech and Adaptation

The deployment of I-Vectors as a tool for speaker recognition/verification[122], language detection[123], accent detection[173], and speaker age[124] showed the growth and trust the speech community put into the algorithm. I-Vectors were developed in 2011 at the Centre de Recherche d'Informatique de Montreal (CRIM) by Dehak, Kenny et al[166]. Prior to this work the group at CRIM developed JFA for use with speech data to address speaker and session variability[174]. Fundamentally, I-Vectors were a natural extension of JFA, but proved to be very effective as a feature preprocessing technique and their own classifier when paired with a simple metric like CD [175, 120].

Fundamentally, evaluations of other ML algorithms relied on tracking the sensitivity and specificity of each experiment and I-Vectors were no different. In fact, they performed inline with other approaches achieving over 90% sensitivity and 90% specificity [171]. Given this development, the approach presented here represents the heart of the technique in as simple a manner as possible. The extensive use of I-Vectors has produced a variety of augmentations, but it would have been unwise to start with a more complex system when transporting the technique to a new field of data. It was decided to minimize as many degrees of freedom as possible while developing I-Vectors for EEGs.

Another problem in adapting this technique was that finding valid speech data was relatively easy. If someone was talking, producing sounds, they were likely producing

valid data. However, that was not the case with EEGs which have a constant stream of data. It was not clear if EEG recordings since background segments are not devoid of information, essentially all data is data of interest. This naturally leads to an increase in background signals in EEGs compared to speech. A sleep study may last for an entire night only to capture a brief 10 minute seizure. Easy for a clinician to correctly identify, but difficult for a ML technique to recognize.

2.4 Machine Learning Algorithms

The breadth of potential algorithms, supervised and unsupervised, was too much to review in depth. Instead, a review of the most notable algorithms referenced in this section and those critical to the validation of I-Vectors were reviewed in the following section. This was meant to provide necessary context to the present field of EEG classification, but was not comprehensive to the rapidly developing realm of ML. Similarly, a brief discussion of FA was included given the frequent use of LDA in supervised and unsupervised techniques and that I-Vectors were predicated on JFA.

2.4.1 Factor Analysis

At a base level I-Vectors reduced the dimensionality of data by finding the most influential features in the given training dataset. In a general sense this is similar to FA which is used to perform blind source separation (BSS), the decomposition of a signal into a linear representation of statistically independent components [172]. While this was the goal, it is difficult to assure linear independence of all the components. As such the techniques are imperfect given the premise of being blind to the true nature of the data.

Two commonly used techniques to achieve BSS are PCA and ICA. From these algorithms more advanced techniques, LDA and QDA, are capable of separating the components of different known classes. They are not able to operate blind, or unsupervised, as they require knowledge of the classes to define class dependent components. Knowing the dependent components they can then resolve the class independent components in an effort to discern the decisions surfaces between the classes. QDA operates in a more generalized space allowing for separation of two or more classes compared to LDA defining separability of a single class from the dataset.

2.4.1.1 Principal Component Analysis

PCA finds the dominant components in a set of data by maximizing the variance of the given features [176]. For a set of data $\boldsymbol{X}$ composed of p columns of features and n rows of observations there exists a vector $\boldsymbol{w}$ capable of maximizing the variance of a given feature.

$$\boldsymbol{V} = \frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \tag{2.4-30}$$

$$\sigma_w^2 = \boldsymbol{w}^T\boldsymbol{V}\boldsymbol{w} \tag{2.4-31}$$

Here $\boldsymbol{V}$ represents the covariance matrix of the data matrix $\boldsymbol{X}$ which is used to find the eigenvectors that become $\boldsymbol{w}$. As eigenvectors are orthogonal to each other, they are each uncorrelated components and produce the p principle components of the $\boldsymbol{X}$.

There are at most n principle components representing unique weightings of the p features. To find the true number of components, q, the number of zero or near zero eigenvalues, $e_z = p - q$, must be found. This linearly independent q-dimensional space represents the true decision surface of the observations. From these operations

it becomes possible to define the critical features and unique observations from the data itself.

2.4.1.2 Independent Component Analysis

ICA separates individual signals from those collected by multiple receivers, commonly known as BSS [172]. The typical example is that of a cocktail party with an equivalent number of microphones and speakers. By using ICA, it is possible to isolate each of the speakers using the data from all of the microphones. This example is referred to as the *Cocktail Party Problem* and exists in many research areas including EEG recordings.

A dataset contains the sequential samples, t, from each recording device and assumes there is a transformation matrix, $\boldsymbol{A}$, that turned the source signals, $\boldsymbol{s}$, into the captured output $\boldsymbol{X}$.

$$\boldsymbol{X} = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} \boldsymbol{A} = \begin{bmatrix} a_1 1 \dots a_1 n \\ \vdots \ddots \vdots \\ a_n 1 \dots a_n n \end{bmatrix} \boldsymbol{s} = \begin{bmatrix} s_1(t) \\ \vdots \\ s_n(t) \end{bmatrix} \tag{2.4-32}$$

$$\boldsymbol{X} = \boldsymbol{AS} \tag{2.4-33}$$

From this output, the features of the recorded signals must be *whitened* before the individual signals can be found. Whitening is a process that transforms the data into a matrix. $\boldsymbol{z}$ that is uncorrelated, but not assured to be independent. The approach is similar to PCA in that it requires eigenvalue decomposition to produce the whitening matrix, V. The matrix $\boldsymbol{E}$ is found from the eigenvectors of $\boldsymbol{X}$ and the diagonal

matrix D contains the associated eigenvalue for each eigenvector.

$$\boldsymbol{z} = \boldsymbol{V}\boldsymbol{x} \tag{2.4-34}$$

$$\boldsymbol{V} = \boldsymbol{E}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{E}^{T} \tag{2.4-35}$$

$$\boldsymbol{z} = \boldsymbol{V}\boldsymbol{A}\boldsymbol{s} = \hat{\boldsymbol{A}}\boldsymbol{s} \tag{2.4-36}$$

Now the transformation matrix, $\hat{\boldsymbol{A}}$, contains only orthonormal components instead of the previous correlated components. This process is necessary as it constrains the solution sets when solving for the independent components.

The *kurtosis* of a signal is one of the many ways to solve for the independent components after whitening. As the kurtosis supports the additive property, it provides a natural process for optimization the non-Gaussian portions of the signal. The expectations, E, of the random variable y's second,variance, and fourth moment are used to find the 'tailedness' of the distribution. With a normalized distribution the expectation of the variance would be 1, but for Gaussian distributions kurtosis would always be zero because the fourth moment is always $3(\mathrm{E}\{y^2\})^2$. This is why the independent components must be non-Gaussian otherwise they cannot be separated out.

$$\begin{gathered}\mathrm{kurtosis}(y) = \mathrm{E}\{y^4\} - 3(\mathrm{E}\{y^2\})^2 \\ \mathrm{kurtosis}(s_1 + s_2) = \mathrm{kurtosis}(s_1) + \mathrm{kurtosis}(s_2) \\ \mathrm{kurtosis}(\alpha s_1) = \alpha^4 \mathrm{kurtosis}(s_1)\end{gathered} \tag{2.4-37}$$

When all the random variables are normalized the variance of y is equal to 1 which bounds the solution by the unit circle. This simplifies the solution to finding a vector that produces the largest amplitude of kurtosis for the given distribution. These

kurtosis based dimensions indicate projections of non-Gaussian distributions which is where the suspected independent signals reside.

$$|\text{kurtosis}(y)| = |q_2^4 \text{kurtosis}(s_1) + q_2^4 \text{kurtosis}(s_2)| \tag{2.4-38}$$

There are other techniques for discerning the projection space of non-Gaussian distributions, Gram-Schmidt, ML estimation, or negentropy, which focus separating independent non-Gaussian distributions. In all instances the mixing matrix $\boldsymbol{A}$ is chosen to be square to simplify the mathematics. The only constrains on the process, regardless of approach, are on the data being statistically independent and that the underlying signals are non-Gaussian distributions. These both require prior knowledge of the signals in the dataset otherwise the results of ICA will be similar to those of PCA, orthogonal uncorrelated feature vectors.

2.4.1.3 Linear Discriminate Analysis

LDA uses the mean and variance of each class in the data to build decision surfaces between the classes. This is achieved by maximizing the distance between the means $\boldsymbol{S}_B$ and minimizing the variances $\boldsymbol{S}_W$ of the features associated with the classes K. Original developed by Ronald Fisher, often called *Fisher's Linear Discriminant*, it seeks to maximize the discriminant factor $J(\boldsymbol{w})$ by finding the vector w [17].

Given two datasets containing n_i observations of each class, a decision surface $\boldsymbol{w}$ can be found.

$$\boldsymbol{X}_1 = \{\boldsymbol{x}_1^1, ..., \boldsymbol{x}_{n_1}^1\} \ , \ \boldsymbol{X}_2 = \{\boldsymbol{x}_1^2, ..., \boldsymbol{x}_{n_2}^2\}$$

$$\boldsymbol{m}_i = \frac{1}{l_i} \sum_{j=1}^{n_i} \boldsymbol{x}_j^i$$

$$\boldsymbol{S}_B = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$$

$$\boldsymbol{S}_W = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_j - \boldsymbol{m}_i)(\boldsymbol{x}_j - \boldsymbol{m}_i)^T$$

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}} \tag{2.4-39}$$

This can be expanded to handle multivarate data by expanding the definitions of $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$. Here $\bar{\boldsymbol{m}}$ represents the mean of the observations n_i across all classes in the training set. Then a sufficient w can be found by maximizing $J(\boldsymbol{w})$ which occurs when $\boldsymbol{w}$ is an eigenvector of $\boldsymbol{S}_W^{-1} \boldsymbol{S}_B$.

$$\boldsymbol{S}_B = \sum_{i=1}^{K} n_i (\boldsymbol{m}_i - \bar{\boldsymbol{m}})(\boldsymbol{m}_i - \bar{\boldsymbol{m}})^T$$

$$\boldsymbol{S}_W = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \boldsymbol{m}_i)(\boldsymbol{x}_{ij} - \boldsymbol{m}_i)^T$$

Classification based off LDA requires an additional step to set thresholds for each class with respect to the resultant eigenvalues produced by $\boldsymbol{w} \cdot \boldsymbol{x}$. Through this metric many approaches can be used to distinguish between the K classes in the multivariate data such as individual or one-versus-all classification.

The multivariate approach often assumes a common global covariance matrix S_X to ensure that $S + W^{-1} S_B$ is diagonalizable. This assures that the eigevenvectors will be caused by the features within the data. To approximate a global covariance

matrix the pooled within-class covariance matrix is scaled by the degrees of freedom between the observations and classes.

$$S_X = (n - K)^{-1} \boldsymbol{S}_W \tag{2.4-40}$$

This results in $K - 1$ eigenvectors as diagonalizablity of a matrix does not ensure unique eigenvectors. In general, LDA is frequently used to perform dimensonality reduction similar to PCA based upon the eigenvalues associated with each eigenvector. Even without reviewing the eigenvalues, LDA always produces one less feature dimension than classes to force discrimination upon the next eigenvector axis.

2.4.2 Algorithms

Numerous algorithms were introduced while reviewing the applications of EEG recordings. The following section highlights the more common algorithms used in ML and those to be compared against I-Vectors. From training datasets the algorithms are able to classify unknown samples by providing a likelihood of a match or a discrete label if given labeled data. These introductions serve only to address the nature of the algorithm, unsupervised or supervised, the process of discrimination, and show the input parameters and type of classification produced.

2.4.2.1 Gaussian Classifiers

Once created, GMMs can be used as the basis for discrimination. As discussed in section 2.3.1.5, the data is broken down into a series of estimated Gaussian distributions. These distributions strive to model classes defined by the data. To identify new data, a likelihood score is generated based upon the distance between each model and the new data sample. Calculating the distance, and thus likelihood, can be done in

a number of ways. Assuming the distributions are Gaussian in nature, the following equation provides the likelihood the point belongs with the model.

Here x is the location in d dimensional space with a known mixture modeled by its mean μ and co-variance Σ.

$$likelihood(x, \mu, \Sigma) = \frac{e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}}{\sqrt{|\Sigma|(2\pi)^d}} \tag{2.4-41}$$

This general form produces the likelihood a sample x could come from a given mixture. The end result becomes a set of likelihoods of the known classes from which to draw a classification label. However, there is no assurance of a data sample exceeding 50% likelihood of any of the classes.

This classifier functions based on the modeled distributions. If the GMMs are created via EM or another clustering method the entire process is unsupervised. However, it is possible make the process supervised by knowing the class means and variances in advance or using labeled data to manual cluster the data. The evaluation of a likelihood based upon a distribution is a fundamental technique used by many ML algorithms. It serves as a natural comparison point for I-Vectors as a preliminary step in their development is to produce GMMs.

2.4.2.2 Naive Bayes Classifier

NBCs make use of probabilities to classify based on discrete conditions. The classifier is built out from Bayes' Theorem which describes the probability of an event occurring given the current conditions. This approach requires knowledge about the events that inform the probabilities making it a supervised algorithm. The two class form of a NBC is

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)} \tag{2.4-42}$$

which provides the likelihood of A given B. In this equation $\mathrm{P}(A)$ and $\mathrm{P}(B)$ represent the independent probabilities of events A and B and the probability of B given A is given as $\mathrm{P}(B|A)$. This expands to multiple conditions T by taking into account the likelihoods of each possible condition with

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B|A_i)\mathrm{P}(A_i)}{\sum_i^T \mathrm{P}(B|A_i)\mathrm{P}(A_i)} \tag{2.4-43}$$

The expansion of the unitary case shows that as the number of conditions increases probabilities for each condition with respect to each class are needed. In a sense the conditions could be features representative of classes or the classes themselves.

The approach is a natural tool for evaluating any modeling technique that produces discrete probabilities assuming they are all independent. Since this cannot always be assumed the technique's performance is dependent on adequate feature selection and class separation. The outcome is a probability of the test event or class occurring that is bounded on $(0\% - 100\%)$.

2.4.2.3 K-Nearest Neighbor Classifier

A KNN classifier uses labeled datasets to assume the class of an unknown sample. This approach is similar to using GMMs, but KNN can only operate with labeled data. Given the k closets neighbors class, the unknown sample is labeled as the highest counted class. The algorithm relies on mapping distances between the data points in their f dimensional feature space [177].

Determining the distance between unique samples provides flexibility in handling non-Gaussian distributions. Unlike GMMs classifiers and similar to NBCs, this algorithm operates directly on the data and not through a model when fed training data. The trade-off becomes having enough data and selecting a sufficient value of

k to produce acceptable classifications. The previous two algorithms relied on the statistics drawn from the training data, but KNN is directly dependent on samples in the training data.

The simplistic nature of and ease of conceptulizing lead KNN to be used in a variety of experiments as a comparative benchmark [14, 152].

2.4.2.4 Support Vector Machines

Another kernel based classifier, SVMs, creates a hyperplane between a target class and all other data. The use of a kernel allows linear and non-linear decision surfaces to be transformed onto a hyperplane for discrimination. This hyperplane maximizes the distance between a target cluster and a non-target cluster [178]. Development of the technique stemmed from considering two normal distributions $\boldsymbol{N}_1 : m_1, \Sigma_1$ & $\boldsymbol{N}_2 : m_2, \Sigma_2$ and an target location x.

$$F_{sq}(x) = \text{sign}\Big[\frac{1}{2}(x-m_1)^T\Sigma_1^{-1}(x-m_1) - \frac{1}{2}(x-m_2)^T\Sigma_2^{-1}(x-m_2) + \ln\frac{|\Sigma 2|}{|\Sigma_1|}\Big] \quad \text{(2.4-44)}$$

In this case $F_{sq}(x)$ resolves to a positive sign indicative point x is in $\boldsymbol{N}_1$ and a negative sign for $\boldsymbol{N}_2$. From this initial equation may variations developed to address non-normal distributions and how to simplify the equation by approximating $\Sigma_1 \approx \Sigma_2$.

Results of SVMs are a binary one-versus-all classification. This provides no way to produce clusters of data nor known the strength of the classifications. As with the other classifiers it builds the hyperplane used for separation from a labeled training set, making it a supervised classifier. As it seeks to maximize the space between clusters additional data is most beneficial when it represents boundary conditions of each class. It has been used on I-Vectors in the speech community [179] and numerous EEG classification tasks [14, 102, 66].

2.4.2.5 Dirichlet Process

A Dircihlet Processes (DP) allows for distributions of distributions to be built in an unsupervised manner. The process produces random variables G_K as sub-distributions from the full dataset's distribution G_0 given a concentration parameter α. In this manner an unlimited number of distributions can be produced from a closed dataset containing $T_1...T_K$ partitions[17] of the data Θ[180].

$$G \approx \mathrm{DP}(\alpha, G_0) \tag{2.4-45}$$

$$\Big[G(T_1), ..., G(T_K)\Big] \approx \mathrm{Dir}(\alpha G_0(T_1), ..., \alpha G_0(T_K)) \tag{2.4-46}$$

Generating new distributions in this manner assures that the average distribution properties are maintained. Those distributions with large α will contribute more heavily, but have a greater likelihood of exemplifying the full dataset's true distribution. Through iterative measures it is possible to produce distributions that separate into naturally defined classes based on the dataset alone.

The clustering of the data occurs via the atoms at each level. An atom is a model of the statistical patterns of some phenomena in the data. At the lowest clustering level only atoms relevant to that level are present, but the next highest level contains these atoms plus their own atoms. Building up towards the highest clustering level means collecting all the atoms along the way. By sharing the atoms across the dataset, it becomes possible to then map similarities based upon the mixture of these atoms at each level [181].

The version used in Wulsin et al.[51], Heirarchical Dirichlet Process (HDP), allows distributions to be drawn across multiple levels of the data at once. This exemplifies

[17] A partition of Θ defines a collection of subsets whose union is Θ. A partition is measurable if it is closed under complementation and countable union.

the use case of a DP for clustering data on multiple levels with minimal prior knowledge. Wulsin built clusters at each level of the data (subject, seizure, and channel) so the knowledge was about the structure of the data and not the contents of the data. This is similar to I-Vectors as features are clustered in the GMMs and then the resultant samples are clustered based on the feature models.

2.4.2.6 Artificial Neural Networks

By applying the functional structure of brain neurons, an algorithm that behaves as a NN can be trained to perform non-linear classification. Each node in the network takes in information from the preceding layer, evaluates an equation to determine its state, and then contributes this activation to the ensuring layer. The connections between nodes have their own weights and the number and depth of layers is based upon the needs of the network. The algorithms referenced thus far included DBNs, RBFNNs, multilayer perceptron neural networks (MLPNNs), and MLPNNs represent a small sample of breadth of NNs.

Depending on the type of data and intended classification goal one NN may perform better than another. The trade-offs between the algorithms stem from the characteristics of the data related to the number of classes and any temporal relationships. At the crux of these algorithms is the need for a large diverse amount of labeled data. Like other algorithms, they learn directly through each sample of data which enables them to be non-linear classifiers. The training methodology is driven by reducing the error in the training dataset through adjusting the weights connecting the nodes and the biases of activation in each node. The complexity of the problem to be solved is often matched by the complexity of the NN.

Of interest to the development of I-Vectors is a Long Short-Term Memory Neural Network (LSTMNN) adaptation capable of quantifying the similarity between two

inputs [182]. By training on ranked input vectors, in the case of Mueller et al. [182] sentences, the algorithm can learn to produce a discrete similarity score. This approach is highly dependent on the initialization parameters and the quality and quantity of training data available given the need to operate on variable length input vectors that represent the same classification.

2.4.2.7 X-Vectors

The I-Vector methodology was improved upon while this work was ongoing by research in the speech community that augmented it with a deep neural network (DNN) [183]. This combined system used I-Vectors and embeddings from a feed-forward deep neural network (the x-vectors) to surpass both of their individual speaker verification performances. The premise of the x-vectors was to utilize the post-statistics pooling layers of the DNN to generate feature vectors. The first embedding was taken from the first affine layer after the statistics pooling and second embedding was taken from the affine layer that received the output of a ReLU (rectified linear units) layer driven by the previous embedding. This made the first embedding a linear representation of the speaker's statistics and the second embedding a non-linear representation of the same statistics. The embeddings and I-Vectors are evaluated using the same process of LDA followed by length normalization and probabilistic linear discriminant analysis (PLDA) to produce the classification scores.

The original research group further refined the technique to the point that it surpassed its I-Vector counterparts [184]. These results were promising for advancing speaker recognition on text-independent datasets, where I-Vectors had been the standard classification technique. However, the x-vector approach is a supervised ML algorithm which relied on speaker labels to build the embeddings from the training dataset before generating the embeddings from the test dataset. They are clearly su-

perior to I-Vectors for speaker/subject recognition, but this approach would be reliant on clinical annotations to expand subject recognition. Additionally, the computational requirements of x-vectors is orders of magnitude beyond that of the presented I-Vector technique, as the original work by Snyder et al [183] utilized 4.4 million parameters over all layers of the DNN.

Chapter 3

METHODS

Those who fail to plan, plan to fail.

Attributed to Benjamin Franklin

The application of I-Vectors on EEGs is a novel concept given that I-Vectors were designed for speech processing. Therefore, there is minimal guidance on how to use I-Vectors on EEG data. As indicated in the background, the foundations of the experiments proposed here came from following the development of I-Vectors within the speech community. The two fields are related in terms of their signal analysis goals, and subject and condition discrimination (Research Aim 1), but their optimization processes may be different (Research Aim 2). In both Aims, the desired goals afford insight into the classification process, which in turn is leveraged into insight about the features, datasets, and EEGs themselves.

3.1 Experimental Outline

The ultimate goal of this research is to provide subject and condition discrimination of EEGs. Prior to this work, this goal was not possible using I-Vectors given the lack of a software tools specifically for EEGs. The first experiments provided classification performance showing that I-Vectors met or exceeded performance of equivalent techniques. Providing competitive classification required an understanding of the technique's trade-offs in terms of features, datasets, and parameters. Running experiments to sweep through the features, datasets, and parameters provided operational

thresholds for the datasets, UBMs, and UBMs for using I-Vectors based classification on EEGs.

In this work the experiments are classified as *Algorithm Benchmarks*, *Parameter Sweeps*, and *UBM-TVM Relationship*. The Algorithm Benchmarks addressed Research Aim 1 (RA1) by testing the performance of I-Vectors against benchmark classifiers, specifically Mahalanobis distance and GMM-UBM. The initial comparisons were carried out using parameters borrowed from speech recognition, which then required optimization through the Parameter Sweeps that addressed Research Aim 2 (RA2). Using the optimal classification parameters, the mechansims by which I-Vectors carried out their classification was resolved through analysis of the relationships between the UBMs, TVMs, and feature sets. These UBM-TVM Relationship experiments addressed Research Aim 3 (RA3) and represented the major contribution to understanding EEGs and multi-modal signal analysis.

Each experiment operated on the same fundamental features, datasets, and evaluations as they built upon each other. This chapter details all the components used to build out the experiments. The ensuing three chapters organize present each of the experiments: Chapter 5 - Parameter Sweeps, Chapter 6 - Algorithm Benchmarks, and Chapter 7 - UBM-TVM Relationship.

3.2 Data

Using heterogeneous data is necessary for validating any statistically rigorous method such as I-Vectors, but EEG data is difficult to obtain. Typically, new data is generated as part of research experiments and/or acquired from hospitals, but rarely if ever enters the public domain. This limits innovation to specific combinations of data and techniques. To mitigate this, only the publicly available datasets from

PhysioNet Database[100] and TUH Corpus[18] were used in this work. While not comprehensive in terms of the variety of subjects and conditions used in other studies this collection provided the necessary breadth to validate the goals of this work. These data include EEG from imagined and actual hand, arm, and foot motion, and normal, abnormal, and seizure clinical EEGs from over 600 subjects.

3.2.1 PhysioNet Database

This EEG data comes from the New York State Department of Health's Wadsworth Center [94] and is a component of the PhysioBank archive maintained by MIT's Lab for Computational Physiology[1]. Within the data bank are EEG recordings pertaining to resting states, imagined motion, and motion tasks. The ~~collected~~ data consist of 64 channel EEGs from 109 subjects performing 14 trials: 12 motion and 2 resting calibration (see Figure 3.1). Information about the subjects (age, gender, handedness, etc) is not provided, making subjects and trials the most applicable decision surfaces.

Each 2-minute imagined-motion/motion trial consists of a series of 30 4.1 second tasks. These alternate between rest states and the computer prompted tasks (T1-T4). The tasks consist of opening/closing left or right fist (T1), imagine opening/closing left or right fist (T2), opening/closing both fists or feet (T3), and imagine opening/closing both fists or feet (T4). The two resting state trials, TR1 Eyes Opened (EO) and TR2 Eyes Closed (EC), are one minute recordings of unprompted subject recordings. From this, three dataset

1. **Physio Full** - All fourteen trials (TR01-TR14)
2. **Physio Single** - One trial of each type (TR01-TR06)
3. **Physio Motion** - One of each motion trial (TR03-TR06)

[1] https://www.physionet.org/pn4/eegmmidb/

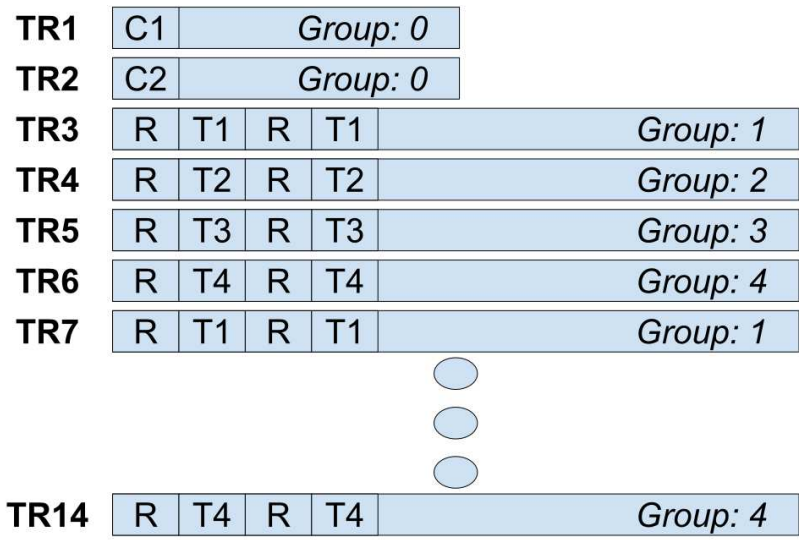


Figure 3.1: The 12 PhysioNet EEG Motor Movement/Imagery Database (PhysioNet Database) trials are broken down by their tasks into 2 resting trials (TR1, TR2) and the four imagined-motion/motion trials (TR3-TR14). This provided depth for subject evaluation, by adding trials, but also for within subject trial evaluation as the repeated trials could be grouped.

These datasets allowed classification experiments on distinct levels of the data. The highest level was subject classification across trials. Beneath that was subject-trial classification, dependent on matching the correct subject and trial. Finally, within-subject trial classification was possible given the grouping of the repeated trials.

The recordings consist of 64 electrodes sampled at 160Hz following a standard 10-20 layout. A 65th channel provides labels for each task during the trials. Since its introduction in 2009, the PhysioNet Database has been used in biometric classifications [105] with respect to task sensitivity [86], subject independence [185], various subject classification schemes [68, 104], and attempts at content based retrieval [186].

3.2.2 TUH Corpus

The Temple University EEG Corpus (TUH Corpus) contains over 25,000 EEGs with their associated medical evaluations. All data comes from patients seen by Temple University Hospital in Philadelphia, Pennsylvania [18]. These recordings represent considerable breadth and depth in terms of patients, medical conditions, and recording conditions. Seizures were the most common diagnosis for patient's with medical records, but stroke and concussion patients are represented as well, while the majority of all recordings are simply indeterminate. In addition to these patients, there are subsets consisting of normal patients and those with indeterminate conditions considered abnormal. These latter classifications (abnormal/normal) along with seizure patients were used to organize 3 distinct datasets:

1. **TUH Normal** - 50 normal patient sessions
2. **TUH Abnormal** - 50 abnormal patient sessions
3. **TUH Seizure** - 411 seizure patient sessions

These datasets allowed for two types of classification experiments. The first was on the subject level, as each was built from unique subjects. The second was developed by combining the datasets to classify them based upon their condition, abnormal/normal/seizure. Further analysis was possible given the associated medical reports, but beyond the time and scope of this research.

Unlike the PhysioNet Database, the TUH Corpus is *in vivo*, leading to a wide array of recording variation. The electrode configurations, sampling rates, and session counts are at the discretion of medical professionals and not a structured research protocol. As addressed in its public release [18], the most common recording configuration consists of 31 electrodes at a 250Hz sample rate. This is substantially fewer

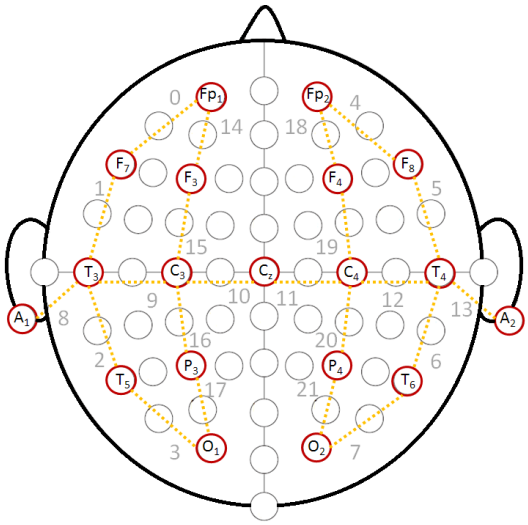


Figure 3.2: The Trans-Cranial Parasagittal (TCP) montage uses a rostral to caudal differential between electrodes to produce channel data. This differential is applied from the ears inward as well to produce 22 distinct channels. Common electrode names are provide with intermediate electrodes left blank. The gray numbers represent the channel index found in the Temple University EEG Corpus (TUH Corpus).

electrodes than the PhysioNet Database, but is enough to produce clinically common EEG montages[2].

3.2.3 Synthetic Dataset

Developing and testing on experimental data alone would make it impossible to provide validation of the software's efficacy; therefore, a synthetic dataset was built. This controlled dataset allowed for two 'ideal' configurations: (1) a dataset with a common feature across all subjects and (2) a dataset with an unique feature for each subject. These datasets were labeled as *simulated*, *static* (simulated with an additional common feature across subjects) , and *unique* (simulated with a unique feature for each subject). Each one contained 10 minutes of data for the simulated 12 subjects and their 22 channels, matching the number of channels in the AutoEEG dataset.

[2]ACNS - Guideline 3: `http://www.acns.org/UserFiles/file/EEGGuideline3Montage.pdf`

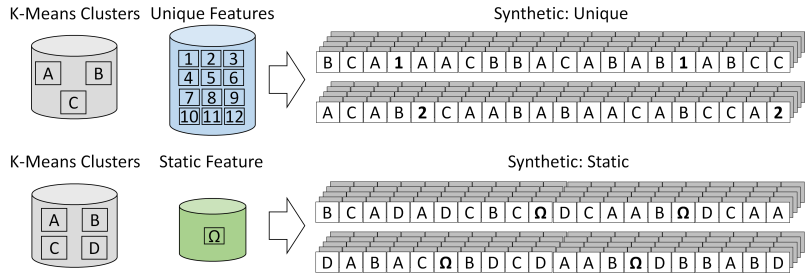


Figure 3.3: The GMMHMM modeled data (gray) and the unique (blue) or static (green) features enable the creation of unique and static synthetic data sets. Only 10% of the simulated data is replaced by the external PhysioNet Database feature. The modeling produced features for each epoch's 22 channels simultaneously to keep the channel-epochs temporal synchronized for each of the 12 simulated TUH Corpus subjects.

Production of the synthetic datasets relied on a Gaussian Mixture Model based Hidden Markov Model (GMMHMM) consisting of 3, 4, or 5 Gaussian models drawn from UBMs. The baseline UBM came from 12 TUH Corpus AutoEEG V1.1.0 subjects using a 16-mixture UBM. The common and unique features came from a single random subject in the PhysioNet Database, also using a 16-mixture UBM. Simulated data contained either 3 or 4 mixtures, allowing the static and unique to add an additional feature containing 4 or 5 mixtures depicted by Figure 3.3.

This produced six unique synthetic data sets: Sim3, Sim4, Sta3, Sta4, Uni3, Uni4, outlined in Figure 3.3. Data was generated for each one-second epoch of each channel as CEP features directly. The distribution of the simulated data followed the weighting of the initial 16 mixture UBM. When the static and unique features were added they overwrote 10% of the simulated data with the new PhysioNet Database-based feature. Authenticity of the raw data was preserved by keeping the synthetic data as similar to the TUH Corpus AutoEEG V1.1.0 dataset as possible, highlighted in Table 3.1.

Table 3.1: Composition of Synthetic Data Sets

Name	Type	Features	Channels	Sampling Rate (Hz)	Duration (s)
AutoEEG	Real	∞	22	100	1200
PhysioNet	Real	∞	64	160	120
Sim3	Simulated	3	22	100	600
Sta3	Static	4	22	100	600
Uni3	Unique	4	22	100	600
Sim4	Simulated	4	22	100	600
Sta4	Static	5	22	100	600
Uni4	Unique	5	22	100	600

3.2.4 Feature Sets

In addition to using multiple datasets, three feature sets were applied to the PhysioNet Database and TUH Corpus: Cepstral Coefficient (CEP), spectral coherence (COH), and Power Spectral Density (PSD). Using multiple feature sets was important because there is no consensus on an optimal feature set for EEGs. PSD features have a long history of use with EEGs [91, 187, 188], as do COH features [64, 162]. CEP are well-established features in the speech processing domain [189, 123]; their application to EEG research was introduced by the Neural Engineering Data Consortium (NEDC) [41].

The COH and PSD features were computed according to the work of LaRocca [64]. The CEP features were built following the standards developed by the speech community [41] and their channels modified to conform with a TCP montage used by neurologists [7]. Thus the feature sets are distinct not only in their mathematical construction, but also their topographical configurations, Table 3.2.

Table 3.2: Feature Set Configurations

Name	Type	Features	Channels
CEP	Original	26	22
	Slim	26	22
PSD	Original	40	56
	Slim	40	19
COH	Original	40	1540
	Slim	40	22

As discussed in the background, EEG recordings can use a variety of electrode configurations. For example, the PhysioNet Database contains 64 electrodes of data, while the TUH Corpus contained a myriad of electrode configurations. Therefore the TUH Corpus set was aligned with the most common standard, the TCP montage, resulting in 19 electrodes organized as 22 differential channels. La Rocca's features consisted of 56 PSD channels and 1540 COH channels making for a larger disparity in channels for each feature set. To address this channel imbalance, the TUH Corpus configuration layout was replicated for the PSD and COH feature sets producing two groups of features. The first was the 55 electrode layouts used by La Rocca [64] and the second time was a mirror of the 19 electrodes from the TUH Corpus TCP montage.

This resulted in a slim feature set consistent of the 22 channel CEP, 19 channel PSD, and 22 channel COH. The CEP and COH confirmed to the TCP layout, but the PSD were not converted to keep them as distinct from the COH features as possible. The benchmark testing against La Rocca's worked used the full feature sets, while all Algorithm Benchmarks and UBM-TVM Relationship experiments used the slim feature sets.

3.2.4.1 Cepstral Features

The CEP-based features were predicated on the success of similar MFCC used in speech recognition. Their adoption for EEG required shifting from a log frequency scale to linear frequency and adjusting the time windows for the Δ and $\Delta\Delta$ differentials. Generation of these features was introduced and detailed by Harati et al in [41], but is outlined here.

The base feature vector consisted of of nine coefficients (seven cepstral coefficients, the frequency domain energy, and the differential energy). The filter banks actually produce eight spectral coefficients covering the following frequency ranges: {0, 1-10, 11-20, 21-30, 31-40, 41-50, 51-60, 71-80 Hz}. However, the zeroth coefficient is discarded and replaced with the frequency domain energy; the differential frequency energy becomes the ninth term. These filters provided a single energy value after bandpass filtering (Hamming) the FFT for each of the listed frequency ranges.

The two energy terms: frequency domain (E_f) and differential frequency energy (E_d) are given as:

$$E_f = \log\Big(\sum_{k=0}^{N-1} |X(k)|^2\Big) \tag{3.2-1}$$

$$E_d = \max\Big(E_f(m)\Big) - \min\Big(E_f(m)\Big) \tag{3.2-2}$$

E_f was derived from the outputs of the filter banks where N are the number of filters and X is the filtered cepstrum frequency output. Using these values within the prescribed 0.9s window of samples, the E_d is found by comparing the maximum and minimum E_f values over the range of m elements in the signal window. These built the first nine features with the remaining 17 coming from the first derivative (Δ) and second derivative ($\Delta\Delta$).

The Δ and $\Delta\Delta$ features used the same equations, but with different window sizes:

$$d_t = \frac{\sum_{n=1}^{N} \left[c_{t+n} - -c_{t-n}\right]}{2\sum_{n=1}^{N} n^2} \tag{3.2-3}$$

Here each sample n in the window N was used to produce a derivative for a given coefficient c centered around time t. Zero padding was used to pad the vector near the beginning and ending of the data. The first derivative Δ used $N = 0.9$. Once resolved, the second derivative $\Delta\Delta$ used the Δ values with a new window of $N = 0.3$. In Harati's work [41], the optimal configuration was found to be a 26 feature vector where the $\Delta\Delta$ for E_d was excluded. This configuration was adopted by the research group and became the consistent feature for the experiments in this work.

3.2.4.2 Power Spectral Density Features

PSD features are derived from the sum of energy over a frequency range for a given time sample. Variation in their creation can be found in their frequency range, number of FFT samples, and filtering of the time signal. The variation of PSD based features used in this work are identical to those of La Rocca et al. [64] which used a frequency range of 0-100Hz, a 100-point FFT, and Hanning windows for filters. The final features were 10-second epochs with 40 PSD values evenly spanning 1-40Hz.

The time series data was filtered with 1 second Hanning window using a 0.5 second overlap. This produced 20 filtered samples for each 10 second epoch centered around each 1 second interval from 0 to 9.5 seconds. These filtered samples were evaluated using Welch's averaged modified periodgram (built into Matlab) with a 100 point FFT to produce 1Hz resolution over the range of 0 to 100Hz. La Rocca's work used the PhysioNet Database data which first had to be resampled from 160Hz to 100Hz prior to the filtering.

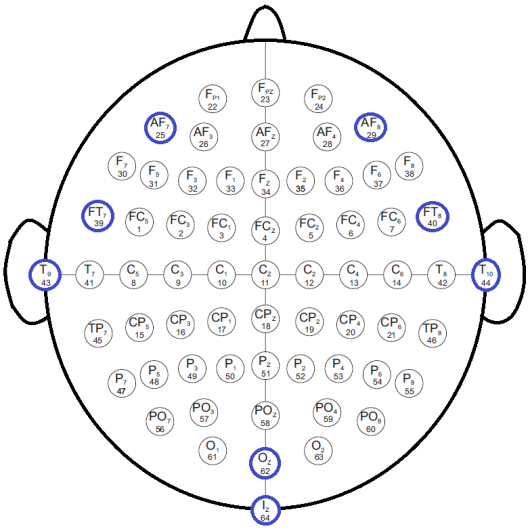


Figure 3.4: The channel layout La Rocca et al used removed 8, highlighted in blue, channels from the overal 64 channel configuration of the PhysioNet Database.

The resultant 100 energy levels were reduced down to only those spanning 1-40Hz. This reduction in frequencies is necessary given (a) the resampling and (b) that the EEG oscillations of interest Delta (0.5-4Hz), Theta (4-7Hz), Alpha (8-14Hz), Beta (15-29Hz), and Gamma (30-40Hz) fall within that range. This resulted in 40 features per EEG channel. The channel count was reduced to 56 from PhysioNet Database's original 64. The discarded channels, highlighted in Figure 3.4, were AF_7, AF_8, FT_7, FT_8, T_9, T_{10}, O_Z, and I_Z.

While originally designed with the PhysioNet Database in mind, these features were readily adapted to the TUH Corpus. Recordings were resampled to 100Hz and pared down to the match the abbreviated 56 channel layout.

3.2.4.3 Spectral Coherence Features

The COH features were proposed by La Rocca as an improvement over PSD features for subject classification. Measuring coherence between electrodes had been used prior

for distinguishing ADHD [139], a general connectivity measure of the brain [34] and auditory oddball paradigms for BCI/P300 responses [82]. Thus they were not novel features, but applied to a broad range of applications beyond subject classification.

These features were generated by quantifying the amount of synchronous energy at each frequency band of each electrode. This was achieved by first building the PSD features and then using them to generate a COH value for each frequency f between two different electrodes i and j, outlined as follows:

$$\mathrm{COH}_{i,j}(f) = \frac{|S_{i,j}(f)|^2}{S_{i,i}(f) \cdot S_{j,j}(f)} \tag{3.2-4}$$

The resultant values were scaled by arctan to normalize their distribution making them bounded on the range $(0, \frac{\pi}{2})$. This configured the final feature set as 1540 'channel' which La Rocca called elements. Each with the 40 distinct frequency bins found through the PSD feature process.

3.2.4.4 Aggregated Datasets

The UBM-TVM Relationship experiments needed subject and condition variation to test classification performance. To achieve, this aggregated datasets were built by combining the PhysioNet Database and TUH Corpus datasets. The combinations of PhysioNet Database's motion data and the TUH Corpus's normal, TUH Corpus's abnormal and normal, or TUH Corpus abnormal, normal, and seizure datasets allowed classification of subjects and known characteristics within a single experiment. This was important to address algorithm robustness and to mitigate any benefits conferred based upon a given dataset-feature-algorithm combination. Each combination was given a designation, Table 3.3 to streamline documentation and discussion.

Table 3.3: Combine Dataset Designations

Designation	Dataset 1	Dataset 2	Dataset 3
`AbnNrm`	TUH Abnormal	TUH Normal	-
`AbnSzr`	TUH Abnormal	TUH Seizure	-
`NrmSzr`	TUH Normal	TUH Seizure	-
`AbnMot`	TUH Abnormal	Physio Motion	-
`NrmMot`	TUH Normal	Physio Motion	-
`SzrMot`	TUH Seizure	Physio Motion	-
`AbnNrmSzr`	TUH Abnormal	TUH Normal	TUH Seizure
`AbnNrmMot`	TUH Abnormal	TUH Normal	Physio Motion
`NrmSzrMot`	TUH Normal	TUH Seizure	Physio Motion
`AbnSzrMot`	TUH Abnormal	TUH Seizure	Physio Motion

3.3 Evaluation Metrics

All experiments were run as subject verification tests. This was inline with La Rocca's experiment which used Correct Recognition Rate (CRR) as their sole evaluation metric. However, given the depth of the datasets and parameter testing to be conducted it was necessary to also include the equal error rate (EER) as well. The performance of I-Vectors has typically been reported in terms of EER, while the EEG research community is typically more broadly focused more on CRR. Exceptions in the literature [86, 103, 163] show results in terms of EER, FAR, FRR, HTER, or Detection Error Tradeoff (DET) curves. For the purposes of this research results were reported in terms of CRR and EER to facilitate readers from both the I-Vector and EEG communities being able to contextualize the experiment performances.

In this work CRR was calculated based on the testing data correctly matching into the enrollment data. The EER was calculated over the entire distance matrix ensuring it evaluated the strength of all matches. This meant if subject 100s second best score was stronger than subject 4's score the EER would be none zero. This is why it was critical to include it for the parameter sweeps, as the CRR masked the majority of the nuance of the full system.

Even with the importance of both metrics, the intended parameter sweeps and comparison points made always displaying both CRR and EER cumbersome and ineffective to the end goal of comparative performance. AS such, the C Metric was defined which combined the CRR and EER by subtracting the EER from the CRR. Thus the threshold for an acceptable C Metric score was set at 0.75 which could represent a CRR of 85% and an EER of 10%. This was primarily used for the expansive Algorithm Benchmarks to showcase performance differences between GMM-UBMs, MD, and I-Vectors.

3.3.1 Mixture Size

For UBMs, TVMs, and I-Vectors the dimension of the underlying mixture model is a critical parameter than can affect performance. Effectively, the n dimensional feature space is modeled by m gaussians; these gaussians are used to train the I-Vectors. As has been the case in the speech community [124, 168, 190], it was necessary to determine the size of the mixture model that would optimize I-Vector performance under different circumstances. While some experiments applied GMM-UBMs previously, their protocols and datasets were not a sufficient starting point[42, 163].

These experiments were used to inform the initial mixture sweep range {2, 4, 8, 16. 32, 64, 128, 256, 512, 1024} used as part of the Parameter Sweeps. After which it was expanded to {2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048} for the Algo-

rithm Benchmarks and UBM-TVM Relationship experiments. The smallest datasets contained 50 subjects, but each dataset had at least 19 channels per subject amounting to a lower bound of 950 distinct subject-channels each with 40 features. From this lower subject-channel bound there the number of epochs in the training and enrollment datastets would change based upon the epoch duration. With the largest epoch duration of 10 seconds, there would be at minimum 9 epochs for each of the 950 subject-channels producing 8,550 unique subject-channel-epochs to model. This value exceeded the upper limits of the two mixture sweeps ensuring overfitting was not a major influence on performance.

3.3.2 TVM Dimensions

The size of the TVM was bounded by the number of mixtures and a depth factor called l from section 2.3.1. As stated, this l value had to be less than or equal to the number of subjects, otherwise models would be built specifically for each subject. Overfitting concerns were addressed with respect to the mixture sizes, but limiting the TVM depth to the number of subjects assured overfitting was impossible in the production of I-Vectors.

This was not strictly required, as the examples used to inform this work would build TVMs with a depth beyond that of the number of subjects [166]. As the TVM is an intermediate step before finalizing the I-Vectors with LDA, the dimension of the TVM could be 1200 for processing data from 75 subjects [191]. However, such options were based on datasets with an order of magnitude more epochs and contained feature vectors double in size than was proposed in this work

Bounding the upper limit was necessary given the dynamic between mixture size, TVM depth, and LDA depth. An upper bound of 200 was chosen because the majority of datasets and aggregated datasets would not exceed 200 subjects. Additionally,

producing the TVM was the most computational intense components of the algorithm requiring a tradeoff of the sweep range and execution time. The lower bound was set at 25, half the smallest subject count. Three incremental values were used to step between the lower and upper bound which resulted in the following sweep range: {25, 50, 75, 100, 200}.

3.3.3 LDA Dimension

The use of LDA to finalize the I-Vectors was well documented by the founders of I-Vectors [121, 166] highlighting their own sweep for optimization with speech data. Thus LDA depth represented a third parameter to consider when building and evaluating I-Vector performance. The upper bound of LDA is determined by the size of its paired TVM. As the range of TVM dimensions was being aligned with the various aggregated dataset subject counts, the LDA dimensions were aligned to operate on a similar scaling.

The lower bound for LDA size was set to 15, slightly less than the TVM lower bound, and the upper bound was set to 100, half the TVM upper bound. Five intermediate values were chosen between the bounds which resulted in the following sweep range: {15, 30, 45, 60, 75, 100}. By focusing on smaller increments this parameter was designed to be less influential than the mixture size and the TVM depth. This sweep range would later be adjusted following the results of the Parameter Sweeps to: {5, 15, 20, 25, 25, 50, 75, 95, 100, 150, 195}.

3.3.4 Epoch Configuration

The final controllable parameters were the number and duration of epochs. Drawing the experiments from the work of La Rocca et al, the initial epoch duration

was 10 seconds with 6 epochs per subject, based around the resting trials of the PhysioNet Database. The epoch durations were expanded to include 5, 2, and 1 second epochs. This naturally altered the number of epochs as the PhysioNet Database contained 1 minute and 2 minute trials which split into a various numbers of epochs for each epoch duration recording combination, show in Table 3.4.

Table 3.4: Number of total epochs per subject as a function of epoch duration.

	Epoch Duration (s)			
Trial Duration (s)	10	5	2	1
60	6	12	30	60
120	12	24	60	120

Based upon reviewer feedback to a prior publication [192] epoch generation was altered to enabled the number of epochs to be independent of epoch duration. This provided another parameter to sweep, number of epochs, which was previously conflated with the epoch duration and trial duration.

3.3.5 Dataset-Feature

Every experiment conducted used all three feature sets, but not every combination of datasets was explored. This was because finding an optimal feature set was beyond the scope of the proposed work. There were not enough available resources in terms of datasets, features, and time to satisfy a robust feature search. However, it was understood that the proposed experiments could offer insight into feature selection which is why every experiment used all three feature sets.

Despite this limitation, using all three feature sets for each experiment provided a comparison point for understanding algorithm-dataset performance. It was hypothesized that one feature set would generally outperform the others, independent of data. Variations in relative performance triggered by mixture size, TVM depth, LDA depth, or epoch settings were used to define areas of interest with respect to these controllable parameters. Additionally, using multiple features mitigated any potential bias generated for stumbling upon an ideal dataset-feature-algorithm combination and being able to identify it as such given the number of dataset-feature-algorithm pairings.

3.4 Implementation

In keeping with the theme of publicly available datasets, the software and hardware solutions were developed to be open sourced. As the research intersects multiple communities it was important that access be given to all regardless of expertise in software development or hardware support. Many of the latest data science solutions required every updating tool kits running on large computing clusters which can limit the use of novel tool kits.

3.4.1 Software

The initial search for I-Vector toolboxes yielded bob.spear[193], Kaldi[194], and Microsoft Research (MSR) Identity Toolbox[195]. The bob.spear toolbox did not work on Windows based machines and Kaldi had proven difficult to implement on the NEDC computing cluster. However, the MSR Identity Toolbox was developed with MATLAB and was easily setup locally and on the computing cluster.

The majority of software was developed specifically for this research with minor components drawn from public sources. A MATLAB toolbox called VOICEBOX[3] was used to support handling of the CEP features generated as Hidden Markov Toolkit (HTK) files. All EDF EEG files were manipulated using edfREAD available through Mathworks MATLAB File Exchange[4].

The decision was made to build using Matlab because it provided a known functional model in the MSR toolbox, would be accessible to both the speech processing and EEG communities, and be robust to hardware/software configurations, and scalable for use on computing clusters. In hindsight there were tradeoffs in terms of performance and flexibility that may have been mitigated by developing the software tools in Python, but the development of this software package was a tertiary goal. Over the duration of the research the Matlab versions started with R2015A and finished on R2017B.

A review of the major facets of the software's workflow is provided in this section. All experiments started with feature creation as the data already existed within the NEDC file system. Once features were produced, a parameter file was written to control the experiment. This file outlined how each process would operate. The experiments were run sequentially to assure each algorithm used the same randomly generated epoch splits for training, enrollment, and testing data. Ultimately, all experiments were run on the NEDC clustering requiring Bash scripts to interface with our Slurm Workload Manager. Those interested in the individual classes and functions should refer to public Git repository's ReadMe[5].

[3] `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`
[4] `https://www.mathworks.com/matlabcentral/fileexchange/31900-edfread`
[5] `https://github.com/izlandman`

3.4.1.1 Feature Creation

The conversion of EEG recordings, stored as EDF files, into CEP, COH, and PSD features was independent of the experiments. This was done to ensure static feature sets and simplified the structure of processing the features during the experiments. Given the number of 'channels' produced from COH features, all feature data was indexed and saved in relation to their epochs.

Thus the number of files produced for each feature set was dependent on subject and number of epochs with channel data organized inside each epoch file. These file lists were the inputs to the experiments where they were aggregated. This tool was written to run with multiple Matlab workers and was supported via a Slurm base script.

3.4.1.2 UBM Class

The use of UBMs was handled through the development of a Universial Background Model (UBM) class in Matlab called *UniBacMod.m*. This simplified the generation, evaluation, and loading/saving of existing models. The generation of the UBMs leveraged Matlab's Single Program Multiple Data (SPMD) parallel computing feature to carry out the EM process on the training data. The enrollment models were built using a parallel MAP adaptation from the generated models and enrollment data. These enrollment models were compared against the testing data to produce log-likelihood ratios which were scored for CRR and EER.

This class controlled the number of UBMs mixtures, the number of EM iterations, and the downsample factor. In addition it held the number of epochs and the resultant UBMs. All of these variables were saved after converting the class to a structure enabling subsequent I-Vector experiments to use the same UBMs.

3.4.1.3 TVM Class

The use of TVMs was handled through the development of a total variability matrix (TVM) class in Matlab called *TotVarMat.m*. This simplified the generation, evaluation, and loading/saving of existing models. Again the EM process used to build the TVM was run using the same parallel processes for the UBM class. The generation of I-Vectors was done in parallel as well, with the option to produce a set of LDA constrained I-Vector in addition to the native TVM I-Vectors. Final evaluations between the I-Vectors were carried out through a parallel cosine distance function to produce the CRR and EER metrics.

The class retained the enrollment and testing I-Vectors and performance metrics binary files, with all other parameters saved as a Matlab structure. Control over the depth of the TVM, depths of the LDA variants, and training steps for the TVM EM were carried out in this class. The constraints previously laid out by the imported UBM are inherited by the TVM class. This assured the number of epochs and UBM parameters were consistent between algorithms. Critically, this allowed the production of I-Vectors from a static UBM produced in a prior experiment.

3.4.1.4 Mahalanobis Evaluation

The use of Mahalanobis Distance as a classifier was borrowed from the work of La Rocca et al. [64]. They developed their experiments using Matlab using the built-in `Mahal` function from the Statistics and Machine Learning Toolbox. Each training/enrollment subject's epochs were used to produce a subject mean. The variances for each feature were drawn from a pooled covariance matrix built from all subject's epoch data.

Evaluation of the the distance matrix between all subjects was used to produce CRRs and EERs aligned to the same epoch, mixture, LDA depth process as the I-Vectors. The resultant distance matrices were saved for each step of the cross-validation process as a binary file. No class was built for this process as it was not the main focus of the proposed research.

3.4.2 Hardware

All of the experiments were run on the NEDC computing cluster, Neuronix. While the cluster supported CPU and GPU parallel processes, the toolkit was written to only support CPU parallelization. Neuronix contained four main identical CPU compute nodes and two minor identical CPU compute nodes. The main nodes consisted of two AMD Opteron 6378s with 16 cores supported by 128GB of DDR3 Ram. The minor nodes consisted of two Intel Xeon E5-2603s with 8 cores supported by 128GB of Ram.

The data server consisted of over 2TB of disk space shared by all the users of NeuroNix.

Chapter 4

NEAR FIELD COMMUNICATION BASED ACCESS CONTROL FOR WIRELESS MEDICAL DEVICES

4.1 Start Here

4.1.1 More Here

4.1.2 And Again

4.2 Restart!

Chapter 5

A PATIENT ACCESS PATTERN BASED ACCESS CONTROL SCHEME

5.1 Start Here

5.1.1 More Here

5.1.2 And Again

5.2 Restart!

Chapter 6

PATIENT INFUSION PATTERN BASED ACCESS CONTROL SCHEMES FOR WIRELESS INSULIN PUMP SYSTEM

6.1 Start Here

6.1.1 More Here

6.1.2 And Again

6.2 Restart!

Chapter 7

BIOMETRICS BASED TWO-LEVEL SECURE ACCESS CONTROL FOR IMPLANTABLE MEDICAL DEVICES DURING EMERGENCIES

7.1 Start Here

7.1.1 More Here

7.1.2 And Again

7.2 Restart!

Chapter 8

CONCLUSION

8.1 Start Here

8.1.1 More Here

8.1.2 And Again

8.2 Restart!

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] O. N. Markand, "Pearls, Perils, and Pitfalls in the Use of the Electroencephalogram," *Semin. Neurol.*, vol. 23, no. 1, pp. 007–046, 2003.

[2] T. W. Picton, "The P300 Wave of the Human Event-Related Potential," *J. Clin. Neurophysiol.*, vol. 9, no. 4, pp. 456–479, oct 1992.

[3] P. Khanna *et al.*, "Modeling distinct sources of neural variability driving neuroprosthetic control," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2016-Octob, pp. 3068–3071, 2016.

[4] Lun-De Liao *et al.*, "Biosensor Technologies for Augmented Brain-Computer Interfaces in the Next Decades," *Proc. IEEE*, vol. 100, no. SPL CONTENT, pp. 1553–1566, may 2012.

[5] B. J. Lance *et al.*, "Brain–Computer Interface Technologies in the Coming Decades," *Proc. IEEE*, vol. 100, no. Special Centennial Issue, pp. 1585–1599, may 2012.

[6] S. Ramgopal *et al.*, "Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy," *Epilepsy Behav.*, vol. 37, pp. 291–307, 2014.

[7] S. Lopez *et al.*, "Automated identification of abnormal adult EEGs," in *2015 IEEE Signal Process. Med. Biol. Symp.*, vol. 37, no. 6. IEEE, dec 2015, pp. 1–5.

[8] H. Nolan, R. Whelan, and R. B. Reilly, "FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection," *J. Neurosci. Methods*, vol. 192, no. 1, pp. 152–162, sep 2010.

[9] E. Schulz *et al.*, "Decoding an individual's sensitivity to pain from the multivariate analysis of EEG data," *Cereb. Cortex*, vol. 22, no. 5, pp. 1118–1123, 2012.

[10] N. Kannathal, M. L. Choo, U. R. Acharya, and P. K. Sadasivan, "Entropies for detection of epilepsy in EEG," *Comput. Methods Programs Biomed.*, vol. 80, no. 3, pp. 187–194, 2005.

[11] V. Lawhern, D. Slayback, D. Wu, and M. Kass, "Efficient Labeling of EEG Signal Artifacts Using Active Learning," *Proc. - 2015 IEEE Int. Conf. Syst. Man, Cybern. SMC 2015*, pp. 3217–3222, 2016.

[12] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms." *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 355–62, feb 2011.

[13] H. Chu, C. K. Chung, W. Jeong, and K.-H. Cho, "Predicting epileptic seizures from scalp EEG based on attractor state analysis," *Comput. Methods Programs Biomed.*, vol. 143, pp. 75–87, may 2017.

[14] D. F. Wulsin *et al.*, "Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement." *J. Neural Eng.*, vol. 8, no. 3, p. 036015, jun 2011.

[15] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz, "Machine-Learning-Based Coadaptive Calibration for Brain-Computer Interfaces," *Neural Comput.*, vol. 23, no. 3, pp. 791–816, 2011.

[16] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Deep Feature Learning for EEG Recordings," *Arxiv*, pp. 1–24, 2015.

[17] A. J. Izenman, *Modern Multivariate Statistical Techniques*, ser. Springer Texts in Statistics. New York, NY: Springer New York, 2008.

[18] I. Obeid and J. Picone, "The Temple University Hospital EEG Data Corpus." *Front. Neurosci.*, vol. 10, no. MAY, p. 196, may 2016.

[19] P. W. Kaplan and S. R. Benbadis, "How to write an EEG report: Dos and don'ts," *Neurology*, vol. 80, no. Issue 1, Supplement 1, pp. S43–S46, jan 2013.

[20] K. M. Tsiouris *et al.*, "An unsupervised methodology for the detection of epileptic seizures in long-term EEG signals," in *2015 IEEE 15th Int. Conf. Bioinforma. Bioeng.* IEEE, nov 2015, pp. 1–4.

[21] A. C. Grant *et al.*, "EEG interpretation reliability and interpreter confidence: A large single-center study," *Epilepsy Behav.*, vol. 32, pp. 102–107, mar 2014.

[22] N. Gaspard *et al.*, "Interrater agreement for Critical Care EEG Terminology," *Epilepsia*, vol. 55, no. 9, pp. 1366–1373, sep 2014.

[23] J. Halford *et al.*, "Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings," *Clin. Neurophysiol.*, vol. 126, no. 9, pp. 1661–1669, sep 2015.

[24] S. C. Warby *et al.*, "Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods." *Nat. Methods*, vol. 11, no. 4, pp. 385–92, 2014.

[25] S. Ghosh-Dastidar, H. Adeli, and N. Dadmehr, "Mixed-Band Wavelet-Chaos-Neural Network Methodology for Epilepsy and Epileptic Seizure Detection," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 9, pp. 1545–1551, sep 2007.

[26] J. J. Halford *et al.*, "Characteristics of EEG interpreters associated with higher interrater agreement." *J. Clin. Neurophysiol.*, vol. 34, no. 2, pp. 168–173, 2017.

[27] C. M. Epstein, "Guideline 7: Guidelines for Writing EEG Reports," *J. Clin. Neurophysiol.*, vol. 23, no. 2, pp. 118–121, apr 2006.

[28] T. Banaschewski and D. Brandeis, "Annotation: What electrical brain activity tells us about brain function that other techniques cannot tell us - A child psychiatric perspective," *J. Child Psychol. Psychiatry Allied Discip.*, vol. 48, no. 5, pp. 415–435, 2007.

[29] E. Westhall *et al.*, "Interrater variability of EEG interpretation in comatose cardiac arrest patients," *Clin. Neurophysiol.*, vol. 126, no. 12, pp. 2397–2404, dec 2015.

[30] K. Gwet, "Kappa Statistic is not satisfactory for assessing the extent of agreement between raters," *Stat. Methods Inter-Rater Reliab. Assessmen*, no. 1, pp. 1–5, 2002.

[31] P. A. Gerber *et al.*, "Interobserver Agreement in the Interpretation of EEG Patterns in Critically Ill Adults," *J. Clin. Neurophysiol.*, vol. 25, no. 5, pp. 241–249, oct 2008.

[32] Z. Z. Wang *et al.*, "Cross-subject workload classification with a hierarchical Bayes model," *Neuroimage*, vol. 59, no. 1, pp. 64–69, jan 2012.

[33] D. La Rocca, P. Campisi, and G. Scarano, "EEG Biometrics for Individual Recognition in Resting State with Closed Eyes," *Int. Conf. Biometrics Spec. Interes. Gr.*, no. Figure 1, pp. 1–12, 2012.

[34] S. Makeig *et al.*, "Evolving signal processing for brain-computer interfaces," in *Proc. IEEE*, vol. 100, no. SPL CONTENT, aug 2012, pp. 1567–1584.

[35] T. Schluter and S. Conrad, "An Approach for Automatic Sleep Stage Scoring and Apnea-Hypopnea Detection," in *2010 IEEE Int. Conf. Data Min.*, vol. 6, no. 2. IEEE, dec 2010, pp. 1007–1012.

[36] A. R. Clarke, R. J. Barry, R. McCarthy, and M. Selikowitz, "Age and sex effects in the EEG: Differences in two subtypes of attention-deficit/hyperactivity disorder," *Clin. Neurophysiol.*, vol. 112, no. 5, pp. 815–26, may 2001.

[37] H. Begleiter and B. Porjesz, "Genetics of human brain oscillations," *Int. J. Psychophysiol.*, vol. 60, no. 2, pp. 162–171, 2006.

[38] J. A. Urigüen and B. Garcia-Zapirain, "EEG artifact removal - State-of-the-art and guidelines," *J. Neural Eng.*, vol. 12, no. 3, 2015.

[39] Q. Gui, Z. Jin, and W. Xu, "Exploring EEG-based biometrics for user identification and authentication," *2014 IEEE Signal Process. Med. Biol. Symp. IEEE SPMB 2014 - Proc.*, 2015.

[40] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, no. 2, pp. 229–240, 2011.

[41] A. Harati *et al.*, "Improved EEG event classification using differential energy," in *2015 IEEE Signal Process. Med. Biol. Symp. - Proc.*, no. December 2015. IEEE, dec 2016, pp. 1–4.

[42] J. L. Marcano, M. A. Bell, and A. L. Beex, "Classification of ADHD and non-ADHD subjects using a universal background model," *Biomed. Signal Process. Control*, vol. 39, pp. 204–212, 2018.

[43] U. R. Acharya *et al.*, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Comput. Biol. Med.*, vol. 100, no. July 2017, pp. 270–278, 2018.

[44] T. Rakthanmanon *et al.*, "Searching and mining trillions of time series subsequences under dynamic time warping," *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 262–270, 2012.

[45] P. Campisi, D. La Rocca, and D. L. Rocca, "Brain waves for automatic biometric-based user recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 5, pp. 782–800, 2014.

[46] B. Porjesz *et al.*, "The utility of neurophysiological markers in the study of alcoholism," *Clin. Neurophysiol.*, vol. 116, no. 5, pp. 993–1018, 2005.

[47] J. A. Coan and J. J. Allen, "Frontal EEG asymmetry as a moderator and mediator of emotion," *Biol. Psychol.*, vol. 67, no. 1-2, pp. 7–49, 2004.

[48] M. Schultze-Kraft *et al.*, "Unsupervised classification of operator workload from brain signals." *J. Neural Eng.*, vol. 13, no. 3, p. 036008, jun 2016.

[49] A. B. Gardner *et al.*, "Human and automated detection of high-frequency oscillations in clinical intracranial EEG recordings," *Clin. Neurophysiol.*, vol. 118, no. 5, pp. 1134–1143, may 2007.

[50] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," *Expert Syst. Appl.*, vol. 32, no. 4, pp. 1084–1093, 2007.

[51] D. F. Wulsin, S. Jensen, and B. Litt, "A Hierarchical Dirichlet Process Model with Multiple Levels of Clustering for Human EEG Seizure Modeling," *Proc. 29th Int. Conf. Mach. Learn.*, pp. 57–64, 2012.

[52] J. G. Bogaarts *et al.*, "Optimal training dataset composition for SVM-based, age-independent, automated epileptic seizure detection," *Med. Biol. Eng. Comput.*, vol. 54, no. 8, pp. 1285–1293, aug 2016.

[53] J. A. Blanco *et al.*, "Data mining neocortical high-frequency oscillations in epilepsy and controls," *Brain*, vol. 134, no. 10, pp. 2948–2959, oct 2011.

[54] V. Bajaj and R. Pachori, "Classification of seizure and nonseizure EEG signals using empirical mode decomposition," *Inf. Technol. Biomed. . . .*, vol. 16, no. 6, pp. 1135–1142, 2012.

[55] W. O. Tatum and W. O. Tatum, IV, *Handbook of EEG Interpretation*, 2nd ed. New York: Demos Medical, 2014.

[56] J. Buckelmüller, H.-P. Landolt, H. H. Stassen, and P. Achermann, "Trait-like individual differences in the human sleep electroencephalogram," *Neuroscience*, vol. 138, pp. 351–356, 2006.

[57] S. L. Wendt *et al.*, "Validation of a novel automatic sleep spindle detector with high performance during sleep in middle aged subjects," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 4250–4253, 2012.

[58] J. Zygierewicz *et al.*, "High resolution study of sleep spindles." *Clin. Neurophysiol.*, vol. 110, no. 12, pp. 2136–2147, 1999.

[59] M. Del Pozo-Banos, J. B. Alonso, J. R. Ticay-Rivas, and C. M. Travieso, "Electroencephalogram subject identification: A review," *Expert Syst. Appl.*, vol. 41, no. 15, pp. 6537–6554, 2014.

[60] P. Tangkraingkij, C. Lursinsap, S. Sanguansintukul, and T. Desudchit, "Personal identification by EEG using ICA and neural network," *Comput. Sci. Its Appl. 2010*, pp. 419–430, 2010.

[61] H. H. Stassen, D. T. Lykken, P. Propping, and G. Bomben, "Genetic determination of the human EEG. Survey of recent results on twins reared together and apart." *Hum. Genet.*, vol. 80, no. 2, pp. 165–76, 1988.

[62] M. Doppelmayr, W. Klimesch, T. Pachinger, and B. Ripper, "Individual differences in brain dynamics: important implications for the calculation of event-related band power." *Biol. Cybern.*, vol. 79, no. 1, pp. 49–57, 1998.

[63] C. E. M. Van Beijsterveldt and G. C. M. Van Baal, "Twin and family studies of the human electroencephalogram: A review and a meta-analysis," *Biol. Psychol.*, vol. 61, no. 1-2, pp. 111–138, 2002.

[64] D. La Rocca *et al.*, "Human brain distinctiveness based on EEG spectral coherence connectivity," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 9, pp. 2406–2412, 2014.

[65] D. La Rocca, P. Campisi, and J. Sole-Casals, "EEG based user recognition using BUMP modelling," *Biometrics Spec. Interes. Gr. (BIOSIG), 2013 Int. Conf.*, pp. 1–12, 2013.

[66] K. Brigham and B. V. K. V. Kumar, "Subject identification from electroencephalogram (EEG) signals during imagined speech," in *2010 Fourth IEEE Int. Conf. Biometrics Theory, Appl. Syst.* IEEE, sep 2010, pp. 1–8.

[67] L. De Gennaro *et al.*, "The electroencephalographic fingerprint of sleep is genetically determined: A twin study," *Ann. Neurol.*, vol. 64, no. 4, pp. 455–460, 2008.

[68] M. Fraschini *et al.*, "An EEG-based biometric system using eigenvector centrality in resting state brain networks," *IEEE Signal Process. Lett.*, vol. 22, no. 6, pp. 666–670, 2015.

[69] M. Näpflin, M. Wildi, and J. Sarnthein, "Test-retest reliability of resting EEG spectra validates a statistical signature of persons," *Clin. Neurophysiol.*, vol. 118, no. 11, pp. 2519–2524, 2007.

[70] A. B. Ajiboye *et al.*, "Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration," *Lancet*, vol. 389, no. 10081, pp. 1821–1830, 2017.

[71] C. Gouy-Pailler *et al.*, "Nonstationary Brain Source Separation for Multiclass Motor Imagery," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 2, pp. 469–478, feb 2010.

[72] P.-J. Kindermans, M. Tangermann, K.-R. Müller, and B. Schrauwen, "Integrating dynamic stopping, transfer learning and language models in an adaptive zero-training ERP speller," *J. Neural Eng.*, vol. 11, no. 3, p. 035005, jun 2014.

[73] B. Blankertz *et al.*, "Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing," *Adv. Neural Inf. Process. Syst.*, pp. 1–8, 2007.

[74] F. Lotte and C. Guan, "Learning from other subjects helps reducing Brain-Computer Interface calibration time," in *2010 IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, no. 2. IEEE, 2010, pp. 614–617.

[75] P.-J. Kindermans *et al.*, "True zero-training brain-computer interfacing–an online study." *PLoS One*, vol. 9, no. 7, p. e102504, jul 2014.

[76] L. A. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalogr. Clin. Neurophysiol.*, vol. 70, no. 6, pp. 510–523, 1988.

[77] N. Karamzadeh *et al.*, "Capturing dynamic patterns of task-based functional connectivity with EEG," *Neuroimage*, vol. 66, pp. 311–317, 2013.

[78] J. Jeong, "EEG dynamics in patients with Alzheimer's disease," *Clin. Neurophysiol.*, vol. 115, no. 7, pp. 1490–1505, 2004.

[79] E. Ba←(s) ar and B. Güntekin, "A review of brain oscillations in cognitive disorders and the role of neurotransmitters," *Brain Res.*, vol. 1235, pp. 172–193, 2008.

[80] S. J. Lupien *et al.*, "The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition," *Brain Cogn.*, vol. 65, no. 3, pp. 209–237, 2007.

[81] A. R. Hassan and M. I. H. Bhuiyan, "A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features," *J. Neurosci. Methods*, vol. 271, pp. 107–118, 2016.

[82] B. Güntekin, E. Ba, and E. Ba←(s) ar, "Review of evoked and event-related delta responses in the human brain," *Int. J. Psychophysiol.*, vol. 103, pp. 43–52, 2016.

[83] C. Vidaurre *et al.*, "Toward unsupervised adaptation of LDA for brain-computer interfaces." *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 587–97, mar 2011.

[84] R. B. Paranjape, J. Mahovsky, L. Benedicenti, and Z. Koles, "THE ELECTROENCEPHALOGRAM AS A BIOMETRIC," in *Can. Conf. Electr. Comput. Eng.*, vol. 2, 2001, pp. 1363–1366.

[85] R. Palaniappan and D. P. Mandic, "Biometrics from Brain Electrical Activity: A Machine Learning Approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 738–742, 2007.

[86] S. Yang, F. Deravi, and S. Hoque, "Task sensitivity in EEG biometric recognition," *Pattern Anal. Appl.*, pp. 1–13, 2016.

[87] R. Mahajan and B. I. Morshed, "Unsupervised eye blink artifact denoising of EEG data with modified multiscale sample entropy, Kurtosis, and wavelet-ICA." *IEEE J. Biomed. Heal. informatics*, vol. 19, no. 1, pp. 158–65, jan 2015.

[88] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, "Trends in EEG-BCI for daily-life: Requirements for artifact removal," *Biomed. Signal Process. Control*, vol. 31, pp. 407–418, 2017.

[89] K. min Su, W. D. Hairston, and K. Robbins, "EEG-Annotate: Automated identification and labeling of events in continuous signals with applications to EEG," *J. Neurosci. Methods*, vol. 293, pp. 359–374, 2018.

[90] J. Gross, "Analytical methods and experimental approaches for electrophysiological studies of brain oscillations," *J. Neurosci. Methods*, vol. 228, pp. 57–66, may 2014.

[91] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 4, no. 2, pp. R1–R13, jun 2007.

[92] A. Subasi and M. I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8659–8666, 2010.

[93] H. Wang and C.-s. Choy, "Automatic seizure detection using correlation integral with nonlinear adaptive denoising and Kalman filter," in *2016 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* IEEE, aug 2016, pp. 1002–1005.

[94] G. Schalk *et al.*, "BCI2000: a general-purpose brain-computer interface (BCI) system." *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1034–43, jun 2004.

[95] H. Kang and S. Choi, "Bayesian common spatial patterns for multi-subject EEG classification," *Neural Networks*, vol. 57, pp. 39–50, sep 2014.

[96] A. Page *et al.*, "A Flexible Multichannel EEG Feature Extractor and Classifier for Seizure Detection," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 62, no. 2, pp. 109–113, feb 2015.

[97] B. C. Armstrong *et al.*, "Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for ERP biometrics," *Neurocomputing*, vol. 166, pp. 59–67, 2015.

[98] E. Maiorana, D. La Rocca, and P. Campisi, "On the Permanence of EEG Signals for Biometric Recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 1, pp. 163–175, 2016.

[99] R. Bódizs, J. Körmendi, P. Rigó, and A. S. Lázár, "The individual adjustment method of sleep spindle analysis: Methodological improvements and roots in the fingerprint paradigm," *J. Neurosci. Methods*, vol. 178, no. 1, pp. 205–213, 2009.

[100] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals." *Circulation*, vol. 101, no. 23, pp. E215–20, jun 2000.

[101] B. Blankertz *et al.*, "The BCI Competition III: Validating Alternative Approaches to Actual BCI Problems," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 153–159, jun 2006.

[102] M. Radha, G. Garcia-Molina, M. Poel, and G. Tononi, "Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal," *Conf. Proc. ... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.*, vol. 2014, pp. 1876–1880, 2014.

[103] S. Marcel, J. D. R. Millán, and J. d. R. Millan, "Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 743–748, 2007.

[104] D. Rodrigues *et al.*, "EEG-based person identification through Binary Flower Pollination Algorithm," *Expert Syst. Appl.*, vol. 62, pp. 81–90, nov 2016.

[105] M. Delpozo-Banos, C. M. Travieso, C. T. Weidemann, and J. B. Alonso, "EEG biometric identification: A thorough exploration of the time-frequency domain," *J. Neural Eng.*, vol. 12, no. 5, 2015.

[106] C. Vidaurre and B. Blankertz, "Towards a Cure for BCI Illiteracy," *Brain Topogr.*, vol. 23, no. 2, pp. 194–198, jun 2010.

[107] M. Spezialetti, L. Cinque, J. M. R. S. Tavares, and G. Placidi, "Towards EEG-based BCI driven by emotions for addressing BCI-Illiteracy: a meta-analytic review," *Behav. Inf. Technol.*, vol. 37, no. 8, pp. 855–871, aug 2018.

[108] J. Martinez-del Rincon *et al.*, "Non-linear classifiers applied to EEG analysis for epilepsy seizure detection," *Expert Syst. Appl.*, vol. 86, pp. 99–112, 2017.

[109] J. A. Blanco *et al.*, "Unsupervised Classification of High-Frequency Oscillations in Human Neocortical Epilepsy and Control Patients." *J. Neurophysiol.*, vol. 104, no. 5, pp. jn.01 082.2009–, 2010.

[110] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," *J. Neural Eng.*, vol. 15, no. 3, p. 031005, 2018.

[111] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *J. Neural Eng.*, vol. 14, no. 1, 2017.

[112] K. A. I. Aboalayon, H. T. Ocbagabir, and M. Faezipour, "Efficient sleep stage classification based on EEG signals," *2014 IEEE Long Isl. Syst. Appl. Technol. Conf. LISAT 2014*, pp. 1–6, 2014.

[113] Shijian Lu *et al.*, "Unsupervised Brain Computer Interface Based on Inter-subject Information and Online Adaptation," *Neural Syst. Rehabil. Eng. IEEE Trans.*, vol. 17, no. 2, pp. 135–145, apr 2009.

[114] N. Kasabov, K. Dhoble, N. Nuntalid, and G. Indiveri, "Dynamic evolving spiking neural networks for on-line spatio- and spectro-temporal pattern recognition," *Neural Networks*, vol. 41, no. 1995, pp. 188–201, 2013.

[115] M. K. Abdullah, K. S. Subari, J. L. C. Loong, and N. N. Ahmad, "Analysis of effective channel placement for an EEG-based biometric system," *Proc. 2010 IEEE EMBS Conf. Biomed. Eng. Sci. IECBES 2010*, no. December, pp. 303–306, 2010.

[116] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," pp. 257–286, 1989.

[117] P. Kenny *et al.*, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 5, pp. 980–988, jul 2008.

[118] H. Behravan *et al.*, "I-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition," *IEEE/ACM Trans. Speech Lang. Process.*, vol. 24, no. 1, pp. 29–41, 2016.

[119] D. A. Reynolds, "Gaussian Mixture Models," *Encycl. Biometrics*, no. 2, pp. 659–663, 2009.

[120] T. Hasan and J. H. L. Hansen, "A Study on Universal Background Model Training in Speaker Verification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 7, pp. 1890–1899, sep 2011.

[121] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in *INTERSPEECH*, no. August, 2011, pp. 857–860.

[122] P. Kenny, T. Stafylakis, J. Alam, and M. Kockmann, "An I-vector backend for speaker verification," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2015-Janua, 2015, pp. 2307–2311.

[123] H. Li, B. Ma, and K.-A. Lee, "Spoken Language Recognition: From Fundamentals to Practice," *Proc. IEEE*, vol. 101, no. 5, pp. 1136–1159, may 2013.

[124] M. H. Bahari, M. McLaren, H. Van Hamme, and D. A. Van Leeuwen, "Speaker age estimation using i-vectors," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1, sep 2012, pp. 506–509.

[125] N. Kasabov and E. Capecci, "Spiking neural network methodology for modelling, classification and understanding of EEG spatio-temporal data measuring cognitive processes," *Inf. Sci. (Ny).*, vol. 294, pp. 565–575, feb 2015.

[126] M. H. Silber *et al.*, "The visual scoring of sleep in adults," *J. Clin. Sleep Med.*, vol. 3, no. 2, pp. 121–131, 2007.

[127] S. K. Loo and S. L. Smalley, "Preliminary report of familial clustering of EEG measures in ADHD," *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.*, vol. 147, no. 1, pp. 107–109, 2008.

[128] S. J. Segalowitz, D. L. Santesso, and M. K. Jetha, "Electrophysiological changes during adolescence: A review," *Brain Cogn.*, vol. 72, no. 1, pp. 86–100, 2010.

[129] S. Fazli, M. Danóczy, J. Schelldorfer, and K.-R. Müller, "L1-penalized linear mixed-effects models for high dimensional data with application to BCI," *Neuroimage*, vol. 56, no. 4, pp. 2100–2108, 2011.

[130] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems." *Neuroimage*, vol. 34, no. 4, pp. 1600–11, feb 2007.

[131] S.-Y. Cheng and H.-T. Hsu, "Mental Fatigue Measurement Using EEG," in *Risk Manag. Trends*. InTech, jul 2011.

[132] S. Dähne *et al.*, "SPoC: A novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters," *Neuroimage*, vol. 86, pp. 111–122, feb 2014.

[133] K. Gwet, *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*, 3rd ed. Maryland, USA: Advanced Analytics, LLC, 2012.

[134] A. Page, J. Turner, T. Mohsenin, and T. Oates, "Comparing Raw Data and Feature Extraction for Seizure Detection with Deep Learning Methods," *Twenty-Seventh Int. . . .*, pp. 284–287, 2014.

[135] V. Gandhi *et al.*, "Quantum neural network-based EEG filtering for a brain-computer interface." *IEEE Trans. neural networks Learn. Syst.*, vol. 25, no. 2, pp. 278–88, feb 2014.

[136] B. Blankertz *et al.*, "Optimizing Spatial filters for Robust EEG Single-Trial Analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, 2008.

[137] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust.*, vol. 28, no. 4, pp. 357–366, aug 1980.

[138] C. Guger *et al.*, "How many people are able to control a P300-based brain-computer interface (BCI)?" *Neurosci. Lett.*, vol. 462, no. 1, pp. 94–98, 2009.

[139] V. J. Monastra, J. F. Lubar, and M. Linden, "The development of a quantitative electroencephalographic scanning process for attention deficit–hyperactivity disorder: Reliability and validity studies." *Neuropsychology*, vol. 15, no. 1, pp. 136–144, 2001.

[140] M. Scheffer *et al.*, "Early-warning signals for critical transitions." *Nature*, vol. 461, no. September, pp. 53–59, 2009.

[141] N. Martin *et al.*, "Topography of age-related changes in sleep spindles," *Neurobiol. Aging*, vol. 34, no. 2, pp. 468–476, 2013.

[142] F. Ferrarelli *et al.*, "Reduced sleep spindle activity in schizophrenia patients," *Am. J. Psychiatry*, vol. 164, no. 3, pp. 483–492, 2007.

[143] E. J. Wamsley *et al.*, "Reduced sleep spindles and spindle coherence in schizophrenia: Mechanisms of impaired memory consolidation?" *Biol. Psychiatry*, vol. 71, no. 2, pp. 154–161, 2012.

[144] M. Mölle, L. Marshall, S. Gais, and J. Born, "Grouping of spindle activity during slow oscillations in human non-rapid eye movement sleep." *J. Neurosci.*, vol. 22, no. 24, pp. 10 941–10 947, 2002.

[145] C. Huang *et al.*, "Discrimination of Alzheimer's disease and mild cognitive impairment by equivalent EEG sources: a cross-sectional and longitudinal study," *Clin. Neurophysiol.*, vol. 111, no. 11, pp. 1961–1967, nov 2000.

[146] A. Lenartowicz and S. K. Loo, "Use of EEG to Diagnose ADHD," *Curr. Psychiatry Rep.*, vol. 16, no. 11, 2014.

[147] I. Buyck and J. R. Wiersema, "Resting electroencephalogram in attention deficit hyperactivity disorder: developmental course and diagnostic value." *Psychiatry Res.*, vol. 216, no. 3, pp. 391–7, may 2014.

[148] S. K. Loo and S. Makeig, "Clinical Utility of EEG in Attention-Deficit/Hyperactivity Disorder: A Research Update," *Neurotherapeutics*, vol. 9, no. 3, pp. 569–587, jul 2012.

[149] T. P. Jung *et al.*, "Removing electroencephalographic artifacts by blind source separation." *Psychophysiology*, vol. 37, no. 2, pp. 163–78, mar 2000.

[150] A. Delorme, T. J. Sejnowski, and S. Makeig, "Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis," *Neuroimage*, vol. 34, no. 4, pp. 1443–1449, 2007.

[151] P. J. Kindermans *et al.*, "True zero-training brain-computer interfacing - An online study," *PLoS One*, vol. 9, no. 7, 2014.

[152] U. R. Acharya *et al.*, "Automated diagnosis of epileptic EEG using entropies," *Biomed. Signal Process. Control*, vol. 7, no. 4, pp. 401–408, jul 2012.

[153] A. Subasi, "Automatic recognition of alertness level from EEG by using neural network and wavelet coefficients," *Expert Syst. Appl.*, vol. 28, no. 4, pp. 701–711, 2005.

[154] N. Huang *et al.*, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. Lond. A*, vol. 454, pp. 903–995, 1998.

[155] S. M. Pincus, "Approximate entropy as a measure of system complexity." *Proc. Natl. Acad. Sci. U. S. A.*, vol. 88, no. 6, pp. 2297–301, mar 1991.

[156] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy." *Am. J. Physiol. Heart Circ. Physiol.*, vol. 278, no. 6, pp. H2039–49, jun 2000.

[157] C. L. Nikias and A. P. Petropulu, *Higher-order spectra analysis. a nonlinear signal processing framework.* Englewood Cliffs, New Jersey: PTR Prentice Hall, 1993.

[158] A. J. Gabor, R. R. Leach, and F. U. Dowla, "Automated seizure detection using a self-organizing neural network," *Electroencephalogr. Clin. Neurophysiol.*, vol. 99, no. 3, pp. 257–266, 1996.

[159] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[160] H. Van Dis *et al.*, "Individual differences in the human electroencephalogram during quiet wakefulness," *Electroencephalogr. Clin. Neurophysiol.*, vol. 47, no. 1, pp. 87–94, 1979.

[161] H. H. Stassen, "Computerized recognition of persons by EEG spectral patterns," *Electroencephalogr. Clin. Neurophysiol.*, vol. 49, no. 1-2, pp. 190–194, 1980.

[162] M. V. Ruiz-blondet, Z. Jin, and S. Laszlo, "CEREBRE: A Novel Method for Very High Accuracy Event-Related Potential Biometric Identification," *IEEE Trans. Inf. Forensics Secur.*, vol. 6013, no. c, pp. 1–13, jul 2016.

[163] P. Nguyen *et al.*, "EEG-Based person verification using Multi-Sphere SVDD and UBM," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7818 LNAI, no. PART 1, pp. 289–300, 2013.

[164] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and Session Variability in GMM-Based Speaker Verification," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, may 2007.

[165] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digit. Signal Process.*, vol. 10, no. 1-3, pp. 19–41, jan 2000.

[166] N. Dehak *et al.*, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 4, pp. 788–798, may 2011.

[167] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, jan 2010.

[168] O. Glembek *et al.*, "Simplification and optimization of i-vector extraction," in *2011 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2011, pp. 4516–4519.

[169] R. McClanahan and P. L. De Leon, "Reducing computation in an i-vector speaker recognition system using a tree-structured universal background model," *Speech Commun.*, vol. 66, no. 1, pp. 36–46, 2015.

[170] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 249–252, 2011.

[171] C. S. C. Greenberg *et al.*, "The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge," *Proc. Speak. Lang. Recognit. Work.*, no. June, pp. 224–230, 2014.

[172] A. Hyvärinen, J. Karhunen, and E. Oja, "Independent Component Analysis," *Analysis*, vol. 26, no. 1, p. 481, 2001.

[173] H. Behravan, V. Hautamäki, and T. Kinnunen, "Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish," *Speech Commun.*, vol. 66, pp. 118–129, feb 2015.

[174] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montr. CRIM-06/08-13*, pp. 1–17, 2005.

[175] M. Senoussaoui *et al.*, "An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech," in *Odyssey Speak. Lang. Recognit. Work.*, 2010.

[176] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1-3, pp. 37–52, aug 1987.

[177] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[178] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[179] S. Cumani and P. Laface, "e-vectors: JFA and i-vectors revisited," in *2017 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, mar 2017, pp. 5435–5439.

[180] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," *J. Math. Psychol.*, vol. 56, no. 1, pp. 1–12, 2012.

[181] Y. W. Y. Y. Teh, M. I. M. M. I. Jordan, M. J. M. Beal, and D. M. Blei, "Hierarchical Dirichlet Processes," *J. Am. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, dec 2006.

[182] J. Mueller and A. Thyagarajan, "Siamese Recurrent Architectures for Learning Sentence Similarity," in *Proc. 30th Conf. Artif. Intell. (AAAI 2016)*, no. 2012, 2016, pp. 2786–2792.

[183] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Interspeech 2017*, vol. 52, no. 2. ISCA: ISCA, aug 2017, pp. 999–1003.

[184] D. Snyder *et al.*, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 5329–5333, 2018.

[185] B. Reuderink, J. Farquhar, M. Poel, and A. Nijholt, "A subject-independent brain-computer interface based on smoothed, second-order baselining," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 4600–4604, 2011.

[186] K. Su and K. A. Robbins, "A Framework for Content-based Retrieval of EEG with Applications to Neuroscience and Beyond." *Proc. ... Int. Jt. Conf. Neural Networks. Int. Jt. Conf. Neural Networks*, pp. 1–8, 2013.

[187] A. M. Dymond, R. W. Coger, and E. A. Serafetinides, "Preprocessing by factor analysis of centro-occipital EEG power and asymmetry from three subject groups," *Ann. Biomed. Eng.*, vol. 6, no. 2, pp. 108–116, 1978.

[188] C. Berka *et al.*, "EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks." *Aviat. Space. Environ. Med.*, vol. 78, no. 5 Suppl, pp. B231–44, may 2007.

[189] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," *IEEE Trans. Acoust.*, vol. 29, no. 2, pp. 254–272, 1981.

[190] H. Behravan, V. Hautamäki, and T. Kinnunen, "Foreign Accent Detection from Spoken Finnish Using i-Vectors," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. August, pp. 79–83, 2013.

[191] M. McLaren and D. van Leeuwen, "Improved speaker recognition when using i-vectors from multiple speech sources," *2011 IEEE Int. Conf. Acoust. Speech Signal Process.*, no. 1, pp. 5460–5463, may 2011.

[192] C. Ward and I. Obeid, "Application of identity vectors for EEG classification," *J. Neurosci. Methods*, vol. 311, pp. 338–350, jan 2019.

[193] E. Khoury, L. E. Shafey, and S. Marcel, "Spear: An open source toolbox for speaker recognition based on Bob," in *2014 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2014, pp. 1655–1659.

[194] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 Work. Autom. speech Recognit. Underst.*, Cambridge, 2011.

[195] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1. 0: a matlab toolbox for speaker-recognition research," *Speech Lang. Process. Tech. Comm. Newsl.*, no. 1, pp. 4–7, 2013.

APPENDIX A

AppendixA

A.1 Start Here

A.1.1 More Here

A.1.2 And Again

A.2 Restart!

APPENDIX B

Appendix2

B.1 Start Here

B.1.1 More Here

B.1.2 And Again

B.2 Restart!