APPLICATIONS AND STATISTICAL MODELING OF ELECTROENCEPHALOGRAMS USING IDENTITY VECTORS

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
30th of November in 2017

Examining Committee Members:

Dr. Iyad Obeid, Advisor, Dept. of Electrial and Computer Engineering
Dr. Zoran Obradovic, Dept. of Computer and Information Sciences
Dr. Joseph Picone, Dept. of Electrical and Computer Engineering
Dr. Yimin Zhang, Dept. of Electrical and Computer Engineering

ABSTRACT

Applications and Statistical Modeling of Electroencephalograms using Identity Vectors

by

Christian Radcliffe Ward

Interpreting electroencephalograms (EEGs) is the domain of trained and experienced subset of clinicians (epileptologists and neurologists). Attempts made through statistical modeling and Machine Learning (ML) algorithms have yet equal their human counterparts. This can be attributed inconsistent inter-rater and intra-rater clinician agreement, the complexity of acquiring signals from the brain, and the variation in people and brain disorders. The knowledge and time required to accurately annotate every EEG recording is not possible given a clinician's daily responsibilities.

Many supervised ML algorithms have been developed with the intent of offloading the annotation process from clinicians. Trained on data sourced from clinicians, these algorithms can only hope to mimic the performance of clinicians. The development of unsupervised ML algorithms struggles to overcome the need of building statistical models from large diverse datasets. However, the creation of the Temple University EEG Corpus (TUH Corpus) coupled with the success of Identity Vectors (I-Vectors) in speech recognition suggest an unsupervised approach may be possible.

Adapting I-Vectors for EEGs will provide insight into the statistical nature of the EEG features, datasets, and classifications. Comparing I-Vectors against other ML algorithms will provide an understanding of how I-Vectors classify EEGs. Using discrepancies between the algorithms it should be possible to unlock the properties of I-Vectors, that proved powerful on speech data, for EEG data. This could lead to novel ways of annotating and understanding EEG recordings without the direct supervision of trained clinicians.

ACKNOWLEDGEMENTS

I would like to thank everyone involved.

Optional Dedication page.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

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

CRIM Centre de Recherche d'Informatique de Montreal

CSP common spatial pattern

DBN Deep Belief Network

DP Dircihlet Process

DT Decision Tree

DWT discrete wavelet transform

EC Eigenvector Centrality

ECG electrocardiogram

EDF European Data Format

EEG electroencephalogram

EER equal error rate

EM expectation maximiation

EMD emperical mode decomposition

EMG electromyography

EOG electrooculography

ERP evoked response potential

ET epileptiform transient

ED Euclidean Distance

FA factor analysis

FAR false acceptance rate

FASTER Fully Automated Statistical Thresholding for EEG artifact Rejection

FSM finite-state machine

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

ICA independent component analysis

ICU Intensive Care Unit

iEEG intracranial electroencephalogram

IMF intrinsic mode function

I-Vector Identity Vector

JFA joint factor analysis

KLD Kullback-Leibler Divergence

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

MD Mahalanobis Distance

MCI mild cognitive impairment

ML Machine Learning

MLE maximum likelihood estimation

MLPNN multilayer perceptron neural network

MFCC Mel Frequency Cepstral Coefficient

mMSE modified multiscale sample entropy

MSE mean squared error

NBC Naive Bayes Classifier

NN Neural Network

PCA principal component analysis

PD periodic discharge

PLED periodic lateralized epileptiform discharge

PLI phase lag index

PMean pooled mean

PhysioNet PhysioNet EEG Motor Movement/Imagery Database

PNN Probabilistic Neural Network

PSD Power Spectral Density

QDA Quadratic Discriminant Analysis

RBFNN Radial Basis Functional Neural Network

REM random eye movement

RF Random Forest

RMS Root Mean Squared

SampEn Sample Entropy

SOM self organizing map

SNN Siamese Neural Network

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

WCCN Within Class Covariance Normalization

WPD wavelet packet decomposition

CHAPTER 1

Introduction

"Qui custodiet ipsos custodes?"

Juvenal's Satire VI

Electroencephalography is a tool for studying the brain. In clinical settings neurologists use it to diagnose conditions such as epilepsy and stroke [1]. It is also used to indirectly study neural responses from various stimuli and neural control in applications such as brain-computer interfaces (BCIs) [2]. More recently, the advent of relatively inexpensive commodity-grade EEG headsets [3] has expanded the field to include areas such as gaming, neuro-modulation, and mindfulness training [4].

With the introduction of digital EEG technology, researchers seek to create digital signal processing tools that can identify or predict neural activity [5]. In clinical fields the technology assists neurologists in reviewing long recordings [6], communicating with patients [7], and processing artifacts [8]. The benefits of these digital tools stem from their multidimensional statistical models [9, 10, 11]. Outside the hospital, researchers have been able to advance BCIs[12] and seizure prediction[13] with the help of these tools.

Historically, computer-based EEG interpretation has been moderately effective, despite large quantities of research [14, 15]. One problem in interpretation 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 and store such large EEG datasets.

Modern digital data collection methods, in 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 and therefore rarely cost effective for clinicians to perform at a fine-grained level [20]. Furthermore, there is only moderate agreement, even among well-trained clinicians, on the correct way to annotate simple events such as various types of spike waves [21, 22, 23, 24]. Building *supervised* ML techniques that mimic clinician performance using annotated[1] data lacking strong consensus is difficult [5]. The difficulty increases when building *unsupervised* ML techniques that operate on unlabeled data [20, 25].

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. As the use cases of EEGs grow, they advance beyond what clinicians typically evaluate making them unable to provide sufficient annotations. With these constraints in mind, the goal of this work is to introduce I-Vectors as an unsupervised machine learning method for EEGs.

1.1 The Landscape of Electroencephalograms

Before outlining the aims of this work, a brief background is provided to ensure an understanding of the relationships between EEGs, algorithms and clinicians. We review how algorithms and clinicians are trained for annotation and classification

[1] In the ML community it is more common to call this type of data *labeled*.

highlights their interdependence and individual short-comings. Specific attention is paid to how clinicians, as individuals and groups, produce annotations which fuel the development of new algorithms. Meanwhile, applications of algorithms are expanding beyond the annotation expertise of clinicians, often into areas outside of clinical settings. With the accumulation of larger and more varied datasets it is necessary to re-evaluate the approaches used in annotating and classifying EEG recordings.

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 [2] 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[3]. 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 the practice of medicine. Principle to their practice 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 [26]. The process of annotating EEG recordings is part of the certification process, but the Epilepsy Foundation

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

[3]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`

contends that 'EEG training for clinicians is inadequate'[4]. In spite of this, clinical annotations remain the best tool for assessing the behavior and state of a brain [27].

Despite all their training, clinicians are not without their faults [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]. These disagreements are most apparent when comparing a clinician's performances on different disorders. Their consensus scores for sleep recordings [24] differ from those of seizure recordings [21] and cardiac recordings [28].

Beyond the type of data, pairwise clinician similarity (Cohen's κ statistic) 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) [28]. 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.

1.1.2 Clinical Annotations

The ability to correctly annotate EEGs 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 presumed 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). This makes it difficult to reuse

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

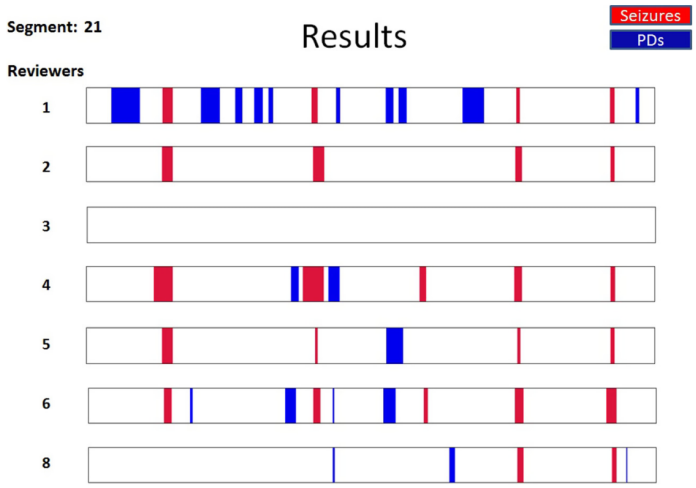

Figure 1.1: In Halford et al. [23], seven reviewers were asked to annotate for seizures and PDs. The annotation results of the hour long recording, Segment 21, show that six reviewers labeled seizure events, five labeled PDs, and one labeled nothing. The quantity of annotation varies as does the spatial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

previously annotated data because it focuses on specific conditions instead of being universal.

To further complicate matters, studies often produce their own datasets because they find existing datasets lack annotations or subject information necessary to address their research questions. This process is especially unforgiving on supervised ML techniques. They are entirely dependent on being trained with a gold standard of annotations, unlike their unsupervised counterparts that do not require annotated data. Thus reliable annotators are a necessity for the majority of EEG based research carried out regardless of the area of interest.

Figure 1.2 illustrates the variance in waveforms that fall under a common annotation label. Not all of these annotations are related to medical conditions, as eye blinks and background are often seen as noise while generalized periodic epileptiform discharges (GPEDs), periodic lateralized epileptiform discharges (PLEDs), spike and sharp wave complexes, and triphasic waves represent the waveforms of interest. The American Clinical Neurophysiology Society (ACNS) defines an exhaustive list of EEG

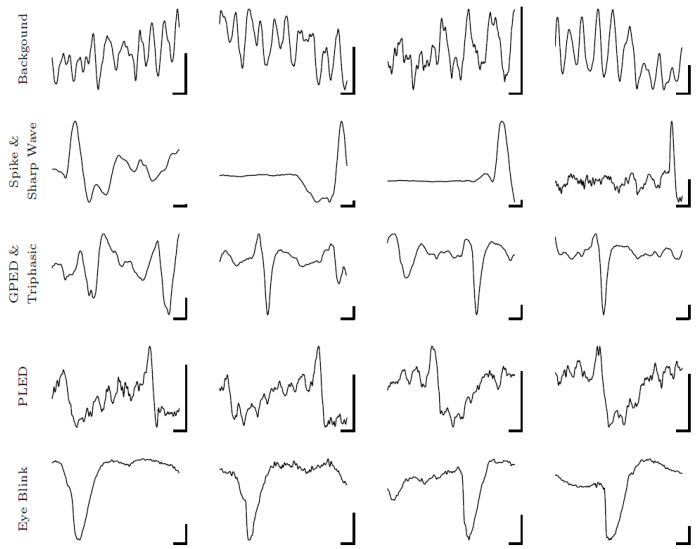

Figure 1.2: Annotations used for the work of Wulsin et al in [14]. Notice the placement of the spike does not need to precede or succeed the sharp wave. GPED and PLED typically occur over a range of channels making them context dependent.

terms which are outlined in [28]. Clinicians are well versed in the terminology, but struggle in their ability to accurately match waveforms with appropriate labels[29].

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 [30], bio-metric identification [31], BCIs [32], evoked response potentials (ERPs) [33], and sleep stage classification [34]. 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, 35] and genetics [36].

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. Without a diverse set of clinician sourced annotations, algorithm based solutions can not progress. This has led to

creation of data specific ML algorithms that struggle to match clinician performance. Instead of universal ML classifiers capable of exceeding clinician performance.

1.1.3 Algorithm Development

The majority of ML algorithms applied to EEG recordings are built to target specific conditions: seizures [14], stages of sleep [34], and evoked responses [33]. There are few attempts to produce general EEG classifiers [37], but pre-processing of EEG data to remove artifacts is widespread [38]. Pre-processing is necessary despite many algorithms operating on clinically recorded data, which highlights the difficulty in dealing with artifacts. A comprehensive solution faces difficulties in terms of the nature of the data (physical activity, stress level, task based trials) and the condition of the subjects (normal, abnormal and awake or asleep).

These single-task solutions suggest the technology is capable, but they fail to increase our ability understand EEGs. A system trained on one dataset is not assured to perform equivalently on another dataset drawn from the same experimental protocol [39]. This often forces algorithms to accept more variance in processing the data to achieve acceptable performance across datasets, such as signals coming from a region of the brain instead of specific electrodes. With this approach, inconsistencies in the origins of signals are mitigated with spatial filters enabling comparisons across the various electrode configurations.

Nearly every EEG condition presents with characteristics primed for dimensionality reduction. Seizure algorithms typically process data in windows on the order of 10s of seconds [40]. Bio-metric algorithms utilize channel subsets to verify a subject [41]. BCIs use spatial filters to target the regions of the motor cortex [42]. ERPs focus on the occipital region where recognition of stimulus is triggered [43]. These techniques are rooted in knowledge gained from the study of EEGs which makes them *domain knowledge*.

ML algorithms are deployed to help us learn about a given datasets, but not all algorithms are equal. In the case where nothing is known about the data, unsupervised ML algorithms must be used to build classifications by modeling its statistical properties. However, in the presence of domain knowledge or *a priori* knowledge, supervised ML algorithms are more effective. Possessing accurate insight into the data helps mitigate bias and error in experimentally collected datasets because less modeling is necessary. The realm of EEGs is no different which is why the ability of clinicians to produce accurate and robust annotations is so important. Supervised algorithms will learn their biases, but unsupervised algorithms struggle to handle depth of EEG data in terms of application and quantity.

The speech processing community developed an unsupervised learning technique, I-Vectors, to produce an unsupervised technique as powerful as supervised one. I-Vectors are able to learn decision surfaces for the accent, age, content, gender, and language of a speaker [44]. Through a series of data modeling utilizing Gaussian Mixutre Models (GMMs) that produce a Universial Background Model (UBM) 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.

1.1.4 Algorithm Applications

BCI, seizure, and sleep data represent a critical set of applications for EEGs, but do not make up the majority of events in the EEGs recordings. An overwhelming amount of any EEG recording is labeled as background. These are signals which clinicians routinely leave unlabeled because they do not contain waveforms relevant to their clinical questions. However, background signals contain information relevant to a subject as documented by the EEG bio-metric community [41]. There are many lesser conditions that manifest in a subject's EEGs recordings such as emotion state

[45], pain [9], and mental focus/workload [46] which a recording annotated for BCI, seizures, or sleep would not contain.

Given the divergence of research interests there are numerous sub-fields of algorithm development pertaining to specific EEG experiments. Some efforts focus on addressing artifacts and noise, while others pick at specific conditions like seizures or ERP, and a few work on generalized analysis of attention/focus and bio-metric classifications. A brief overview of the aforementioned EEG applications is presented, but is not exhaustive. The intent is to help contextualize where EEG research presently stands with respect to applications and tools.

Seizures A substantial portion of work in this field focuses on correctly identifying and locating seizures [47, 48, 49, 50]. By isolating seizure events researchers can focus on the properties of the seizure for the purposes of classification and waveform modeling [14, 51, 52]. The knowledge gained in this process makes it possible to predict seizures in real-time [5, 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 preprocessing for algorithms [54]. 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 (PSD) [55]. While seizures often manifest during sleep, other issues can also be addressed such as sleep apnea[34] and overall brain functionality/health[56].

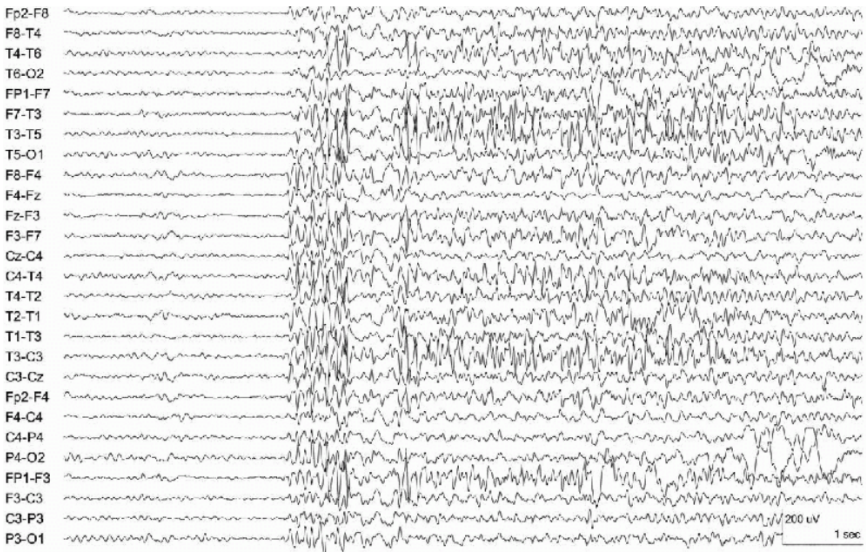

Figure 1.3: A segment of an EEG recording taken from a subject at the onset of a generalized seizure. Note that after the seizure starts, activity is not uniform across all channels. Image sourced from Tatum and Tatum[53].

Bio-metrics Multiple studies have focused on the unique nature of EEGs relative to individuals irrespective of disease and disorder [57]. The results of such work suggest that individuals have distinct EEG fingerprints [58, 59, 60] which also posses inheritable qualities [59, 61]. A major theme in bio-metrics is understanding how different brain states impacts these fingerprints. The work of Rocca et al showcases brain distinctiveness when using a common testing state of resting eyes closed [31], spectral coherence as discrimination feature [62], and techniques to reduce the feature set into sparse mappings [63]. Some approaches overlap with other applications by invoking response potentials [64], focusing on specific brains states of sleep [65], or restful states with eyes open and closed [66]. Even the longitudinal stability of bio-metric EEGs is tested [67] to determine viability for long term applications.

Brain Computer Interfaces BCIs 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 [68, 69]. This requires algorithms robust to changes

Figure 1.4: 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[53].

in subjects, but sensitive to spatial and temporal facets of EEG recordings [43, 70]. Development of subject invariant algorithms has led to disparate training protocols with transfer learning using multi-subject models[71] and zero-calibration training being subject specific [72]. This leads to a similar problem as sleep, where the waveforms are well understood, but their manifestation across populations complicates their performance.

Evoked Response Potentials BCI is a broad term and could possibly include ERPs, but ERPs are a stimulus response and not a voluntary action. A well documented case of ERP is the P300 signal that triggers in the pariatal/occiptal region 300 milliseconds after seeing an image of interest [7]. 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 [73]. This approach allows a brain to communicate without the need

of a body, but also applications for testing processing time of visual and auditory stimulus response [74].

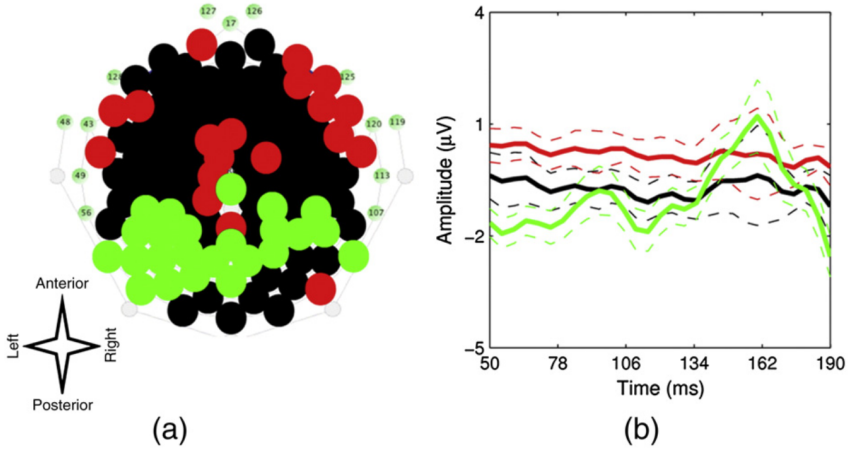

Figure 1.5: A 2D mapping of the electrodes and their group averaged waveform (solid lines). The standard deviations of the channel averaged are given as the dashed lines. Image sourced from Karamzadeh et al.[74].

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 [75], alcoholism [76], and mental disorders such as attention-deficit/hyperactivity disorder (ADHD) and Bi-Polar disorder [77] 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 [78] and a brain's workload during attention dependent tasks [46].

1.2 Research Proposal

A clinician's primarily focus is on treating patients. Asking clinicians to produce perfectly annotated recordings to support algorithm research is not in the best interest of their patients or their productivity. Attempting to use algorithms to drive

advancements in annotation is a circuitous problem when they rely on the annotations from the clinicians. Therefore, if clinicians do not have the time to annotate large sets of data and their annotations are not always in agreement, how will advances be made?

The most direct solution is to find a way to annotate recordings without involving clinicians. As discussed there are ML based solutions, but there is minimal consensus on which approach is best. Despite the success of ML algorithms, fundamental problems exist that they cannot overcome. These include the quality of the recording, enough annotated data of the condition, an acceptable feature set, and consistent channel layouts. The core issue is identifying what characteristics of the EEG are relevant to the task. In most instances some prior knowledge is applied to reduce the dimensionality, and thus the uncertainty, in the ML algorithm. This forces a reliance on the annotations produced by clinicians, despite the quality and quantity of annotations often being are disparate of each other.

Annotations are critically important for supervised ML algorithms. They are presently the dominant ML approach to classifying data which means clinicians control their performance through their own annotations. To alleviate this constraint unsupervised ML algorithms must be developed that are capable of equaling the performance 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 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 is to lay the foundation for an unsupervised ML system that classifies and clusters EEG recordings. Preliminary work indicates it is possible for I-Vectors to perform subject verification and sort data by similarity[5]. These results are promising, but are not enough to justify the adoption of this technique. Little will be gained if it cannot determine what makes such classification and clustering possible.

In addition to understanding how the system operates on EEGs, it is necessary to prove I-Vectors can offer comparable performance to existing standards. In terms of waveform and brain state classification this includes both ML algorithms and clinicians. However, given the 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. Thus clinicians again provide a standard against which algorithms are measured, despite their inability to come to those standards from raw EEG data.

With the goals of classification and clustering via explainable statistical properties inherent in the data, three questions can be posed:

Research Aim 1: How are EEGs statistically differentiated through the modeling process of UBMs, TVMs, and I-Vectors?

Research Aim 2: How does I-Vector classification and clustering compare to other applicable ML techniques?

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?

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

By answering these questions, insight into the nature of I-Vectors and EEGs technology will be advanced. This is because the majority of work with I-Vectors has focused on speech signals which are better understood at the physiological level. The lack of standardization in processing EEG signals compared to speech signals requires that the proposed experiments be conducted to control for the increased dimensonality given the lack of domain knowledge.

1.2.2 The Research Motivation

The reason to carry out this work is the clinical and academic need for annotated EEG recordings. A Catch-22 exists in that the best supervised ML techniques require a strong foundation of knowledge, consisting of labeled data, to produce results on par with trained clinicians. Without a peer-reviewed training dataset algorithms are unable to adequately annotate unlabeled datasets. The algorithms are beholden to the bias of the clinicians. Clinicians are biased by their training and experience, of which will never be comprehensive for all use cases of EEGs.

This lack of universality means ground truth data must be produced for the training of supervised algorithms and clinicians. While clinicians would be the ideal source for this data, their time is better spent with their patients and they lack consistent inter-rater agreement. As supervised ML approaches are reliant on annotated data, their performance suffers when they are trained on data that comes from unreliable sources. This suggests a third approach, unsupervised ML, is needed to support the continued development of supervised ML algorithms and clinicians.

Unsupervised ML is not without its drawbacks, largest among them a dependency on an equitable distribution of data. Fortunately the availability of EEG is the easier to address than the availability of accurately annotated EEG data. While there have always been numerous publicly accessible datasets, the existence of the TUH Corpus suggests there many be enough data available to adequately train an

unsupervised algorithm. It is through these databases and advanced unsupervised ML techniques it is possible to deepen our understanding of EEGs because the approach is mathematical and free from external sources of bias.

Part of expanding our EEG knowledge centers on formalizing how the data should be processed. A substantial amount of research focuses on finding the best feature-algorithm combination for given classification tasks. This is due to the disparate nature of the conditions, subjects, and study protocols. I-Vectors are built on UBMs and TVMs that strive to constrain the relationships between the natural modes of the data while providing dimensionality reduction. This makes I-Vectors an enticing option at encapsulating the disparity between datasets while producing transparent decision surfaces.

From these decision surfaces it should be possible to expand our knowledge of EEGs through the statistically modeling that produces I-Vectors. The ability to simultaneously classify and cluster datasets provides the ability to explore the impact of variations in features and training data. Accepting the imperfect nature of unsupervised ML means that imperfect results are accepted and encouraged. These edge cases should provide insight into the functionality of I-Vectors as they are deployed onto a new type of data. From these insights relationships between EEG data and features should all be possible helping close the loop on EEGs annotations.

1.2.3 The Research Experiments

The Aims of this work will be addressed with three sets of experiments: *Care*, *Principal*, and *Comparison*. Upon completion of the experiments, the process of producing I-Vectors from EEG data should be well understood along with what properties of EEG and I-Vector make this approach viable for producing annotations in an unsupervised manner.

Core Experiments The purpose of the *Core Experiments* is to determine optimal operating parameters for applying I-Vectors to EEGs. This addresses RA 1 by measuring the significance of specific features, channels, UBM mixture sizes, and the TVM training process. Sweeping through each parameter sequentially results in a set of parameters optimized for the specific dataset while providing general trends for each parameter. The statistical decomposition of each dataset and I-Vector development process provides background and baseline results enabling comparisons against the other experiments.

Principal Experiments In order to validate I-Vectors as an option for classification and clustering of EEG data their performance must be compared against a suite of ML algorithms. The algorithms will be evaluated through their sensitivity and specificity and when applicable their ability to cluster. In a sense these experiments mirror the inter-rater and intra-rater clinician evaluations. These *Principal Experiments* address RA 2 through a series of leave one out cross validation (LOOCV) experiments based on subject and channel classifications[6].

Comparative Experiments Datasets built from classification outliers will be used to model the edge cases of each algorithm. The epochs and their associated I-Vectors will be evaluated for the distribution of their features to resolve RA 3. Using the knowledge gained from the previous experiments it should be possible to modify the feature sets and training data to alter the classification results of the I-Vectors and other algorithms. If performance gains are consistent across algorithms it would suggest novel understanding of EEG data has been gained. Otherwise, the shifts in performance would suggest there are facets of EEG data that lend themselves to I-Vector.

[6]There is a need for some labeling as not all the algorithms for comparison operate in an unsupervised manner.

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 [45], interprets stimuli [60], and changes due to diseases and disorders [76]. 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 [33]. This changes the discrimination techniques when dealing with BCI to spatial and temporal features [70, 79]. Auditory and visual stimulus response [7] have distinct spatial patterns as well adding to the diversity of BCI waveform morphology [72].

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, 80], while algorithms operate on epochs of seconds [14, 72].

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 [81]. 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[82]. Adult EEGs also provide insight into numerous conditions such as alcoholism [76], Alzheimer's Disease [75], brain development [83], emotion [45], and stress [78].

In a research setting, BCIs promote a deeper understanding of brain functionality by allowing those with disabilities to communicate [7] and regain functionality [68]. 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 [32], and bio-metrics [84]. 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[3]. 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[38, 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 lay-out, Figure 2.1, based upon relative distances [85]. 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

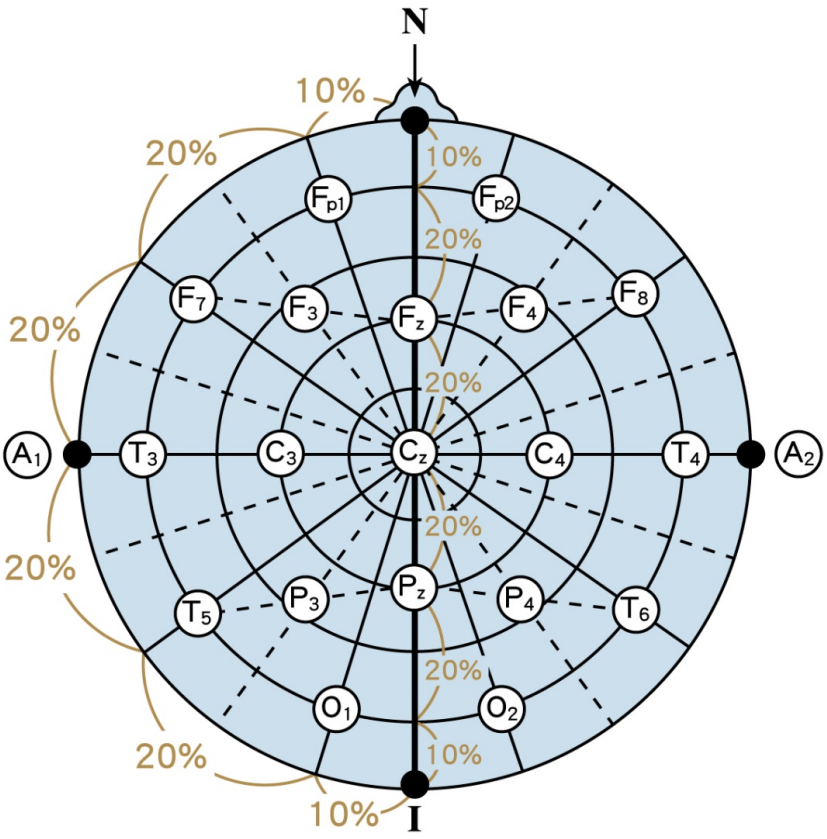

Figure 2.1: The 10-20, 10-10, and 10-5 layouts for EEG electrodes utilize a proportional unit of measure for the distribution of electrodes. The first number represents the distance of the electrodes from the nasion and inion and the second represents the space between subsequent electrodes. With this approach adding electrodes does not change the location of the previous electrodes. Image sourced from [86].

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 [88], while the stages of sleep affect all but the gamma band [34].

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

[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 [87].

[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

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

*When dealing with motor cortex signals it is common to encounter the Mu band (9-11Hz) which resides within the Alpha band.

This work applies to the PhysioNet EEG Motor Movement/Imagery Database (PhysioNet) dataset and the 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 dataset and predominantly seizure focused TUH Corpus dataset.

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.

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,

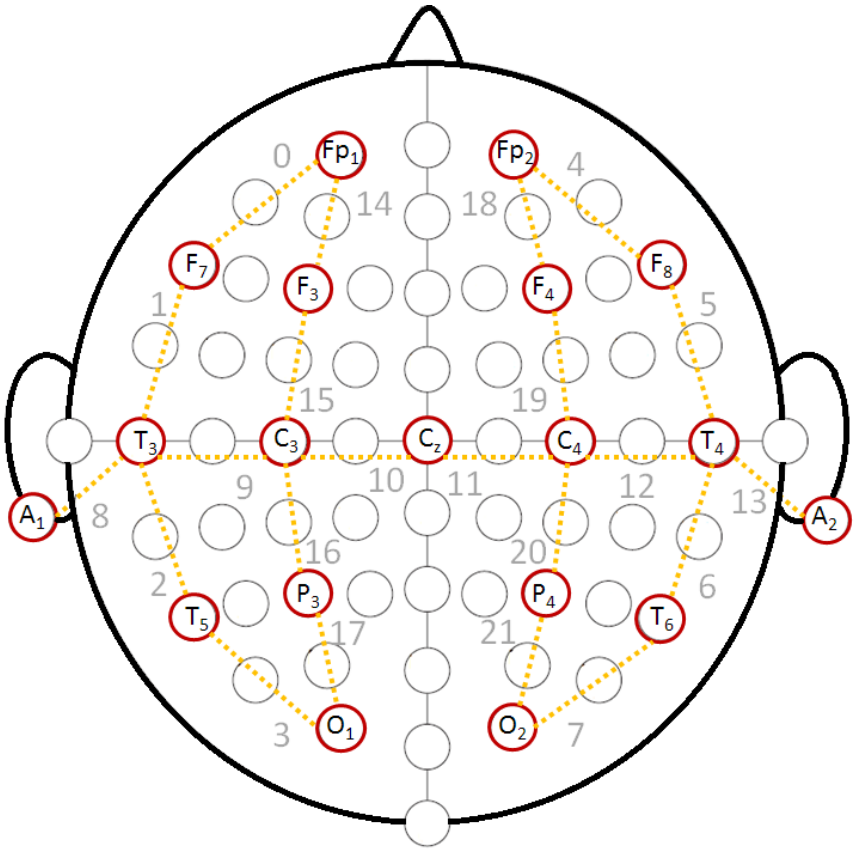

Figure 2.2: The TCP Montage channels (red) used by the TUH EEG corpus is overlaid on the PhysioNet channel layout. Each montage link (orange) is assigned an index for storing the montage channel (gray) data in the corpus. The proper 10-20 channel names (black) are provided for the montage channels.

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.

PhysioNet EEG Motor Movement/Imagery Database

The PhysioNet data contains 109 subjects following computer prompted motion/motion imagery trials at the New York State Department of Health's Wadsworth Center [42]. 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 website [89].

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' 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

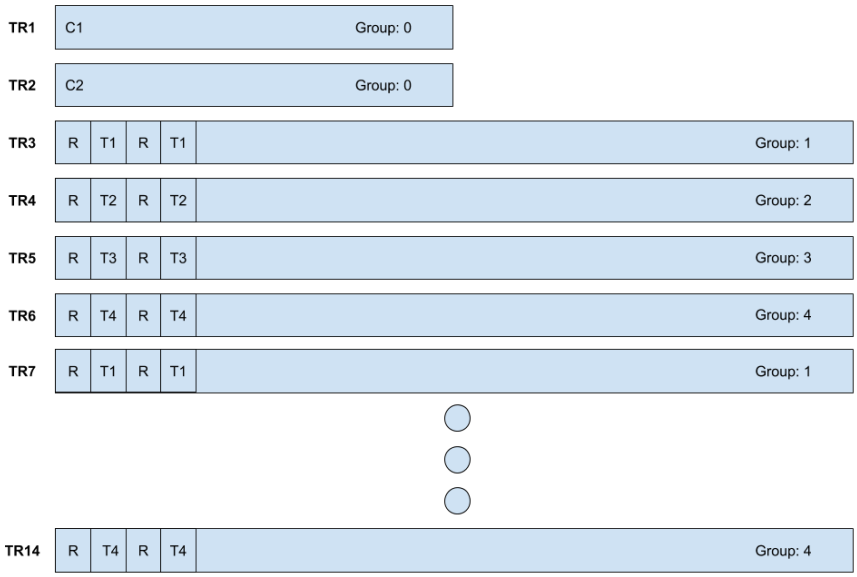

Figure 2.3: Each subject from the PhysioNet data set completed 14 trials. Two of these trials (TR1 and TR2) are one minute calibrations trials of resting eyes open and resting eyes closed. The remaining 12 trials are two minute recordings of a predefined sequence consisting of a task state and resting 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.

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 [26]. These reports must document 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

[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.

clinically relevant interpretation of the patient's condition may not be forthcoming without reviewing the reports of other tests and/or subjects [26]. 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 [90].

In contrast, an algorithm's approach to annotation is much more broad. Depending on the desired outcome, algorithms can perform a normal/abnormal classification [6], annotate specific epochs [14] or combine these approaches to classify EEG recordings [34]. 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 waveforms, epochs, and entire recordings across subjects is important in the development of robust BCI [71] and bio-metric applications[41]. In this area algorithms exceed the ability of clinicians by shifting how EEG recordings are evaluated through novel channel and feature selection [63, 64, 66].

Specifically, bio-metric algorithms can determine the similarity of one subject to another [41, 61]. 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 [91]. This makes their application to EEG recordings

interesting as I-Vectors may be capable of bridging classification between algorithms and clinicians.

Defining Similarity via Cohen's Kappa

It is difficult to produce annotated sets of EEG recordings without clinical support. To ensure the accuracy of these sets it is necessary to have multiple clinicians annotate the same data to build a consensus-based annotation. This process invites each clinician's bias into the annotation process which must be tracked and controlled in terms of intra-rater and inter-rater similarity scores. These scores provide a sense of strength of a clinician's ability and robustness of a dataset as a function of *agreement* evaluated as Cohen's Kappa (κ).

Table 2.3: Table of Cohen's Kappa

		S1	
		A	B
S2	A	q	w
	B	z	x

Given two raters and their tallies for class A or B in Table 2.3, their inter-rater agreement κ is calculated as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{2.2-1}$$

$$p_o = \frac{q + x}{q + w + z + x}$$
$$p_e = \frac{q + w}{q + w + z + x} * \frac{q + z}{q + w + z + x} + \frac{z + x}{q + w + z + x} * \frac{w + x}{q + w + z + x} \tag{2.2-2}$$

In the above equation, p_o finds the percentage of agreement between the two raters[5]. Then p_e finds the percentage the raters chose the same label, how often S1 chose A

[5]In the event the two raters are the same clinician, the agreement represents intra-rater agreement instead of inter-rater agreement.

and S2 chose A. The calculated expectation of similarity, p_e, is used to control for the outcome of similarity, p_o. The grades of agreement are quantified as follows: $\{ < 0$ poor; $0 - 0.20$ slight; $0.21 - 0.40$ fair; $0.41 - 0.60$ moderate; $0.61 - 0.80$ substantial; $0.81 - 1.00$ almost perfect$\}$ [92].

2.2.1 Clinician Classification

When clinicians annotate EEG recordings it is important that they use the same terminology when describing waveforms. Without a shared vocabulary EEG reports would be ineffectual for diagnostics and documentation[26]. Gaspard et al.[22] tested 49 clinicians' agreement on terminology by asking them 409 questions about 37 pre-selected EEG waveforms. This removed the task of finding the epochs of interest which enabled the experiment to isolate a clinician's ability to describe the waveform activity in the epoch.

Each clinician's background varied in terms of experience (2-15+years) and training (adult or pediatric neurology). The epochs were sourced from critical care patients exhibiting PLEDs, GPEDs, seizures, and other rhythmic activity. Each 10 second epoch was a modified biploar montage filtered for signals between 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 2.4.

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 indicate a universal blind spot that would be passed on to an algorithm built from this annotated data. Since the

Table 2.4: 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.

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)

contents of epochs are known, this show how difficult it is for clinicians to agree on labeling of known wavefroms.

The cause of these biases may be that clinicians are evaluated on their annotations indirectly. Their diagnosis is not solely based on the EEG, but also the patient's medical history. In Halford et al. [90] the importance of detecting epileptiform transients (ETs) is critical for diagnosing epilepsy. Failing to annotate some of the ETs does not change the diagnosis because the clinicians are primed to make a decision about epilepsy. Individually the 18 tested clinicians are unable to produce a Gwet agreement coefficient[6] over 0.50 with the rest of the group. This indicates a weak

[6]The Gwet's AC2 is an alternative to κ statistics for quantifying inter-rater similarity, but is bounded over the same range [93].

agreement among the clinicians. Despite varying levels of certification and years of practice, there are no distinct indicators of what produces a better annotator.

The difficulty in producing accurate annotations with respect to others is a mixture of finding the waveforms and then correctly labeling them. These problems are documented to various degrees as clinicians were tested for the annotation skills on critically ill patients [29], patients exhibiting seizures [21, 23], comatose cardiac patients [28], and sleeping subjects [80]. The following two sections review clinician inter-rater and intra-rater agreement as a function of the type of EEG data.

Clinician Inter-rater Agreement

The previous section discussed this broadly and with the benefit of the waveforms being pre-selected. However, when clinicians are asked to annotate longer epochs the discrepancies shift from clinical knowledge to issues of annotation style. Inter-rater agreement is the ability of one clinician to agree with one or more other clinicians.

A pedantic instance of this is seen in Figure 2.4 where two clinicians have labeled a 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 5 other seizure events, but their agreement is weakened because of their three misidentified events.

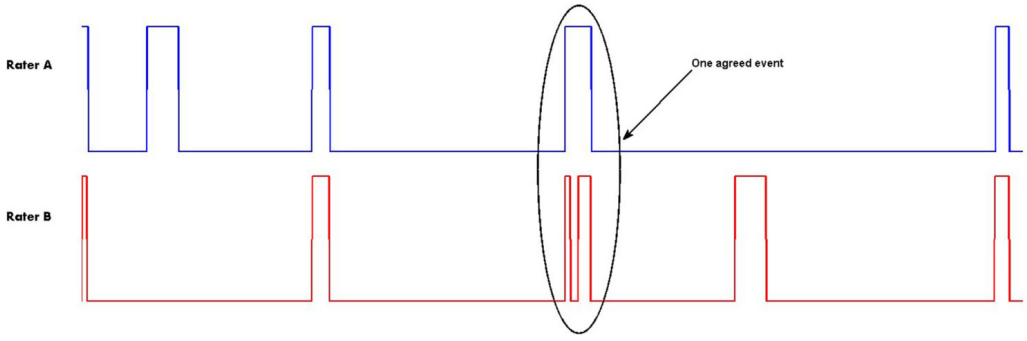

Figure 2.4: An example of how open ended annotation styles lead to inconsistencies in evaluating the accuracy of inter-rater agreements.

This example comes 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 highlight the difficulty in finding consensus by suggesting it goes beyond what their background and experience. There is a clear issue in how clinicians select waveforms in the recordings, which results in less data being included in any gold standard.

Gerber et al.[29] 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[7]. The results, Table 2.5, show the clinicians' consensus is stronger on the shorter epochs (0.04-0.68) than the longer epochs (0.07-0.44).

Table 2.5: 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

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

The most critical labels (rhythmic/periodic vs. excluded, localization, and morphology) exceed 65% agreement, but only rhythmic/periodic exceeds 80%. This means that on average each clinician failed to recognize 20% to 35% of what the other clinicians annotated. Without definitively labeled data it is 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 is possible their performance is impeded by alignment issues similar to those seen in Halford et al.'s work. The results otherwise suggest that the clinicians agree at a moderate to fair level.

Gerber et al.'s at best substantial inter-rater agreement is inline with Halford et al.'s and this trend continues 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.

Westhall et al. [28] asked 4 clinicians 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 2.7. Meanwhile, the background and periodic patterns produced inter-rater κ statistics between -0.07 to 0.82, Table 2.8.

Just as the results of Gerber et al. showed strongest performance for critical waveforms, Westhall et al. does as well. However, performance outside these critical waveforms is extremely poor in terms of classification agreement and κ statistics. This may be caused by the increase in classification categories, compared to Gerber

Table 2.6: 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

Table 2.7: 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)

et al., Grant et al., or Halford et al, but more likely suggests the clinicians fundamentally disagree over the non-prespecified EEG patterns. Where background EEG or periodic patterns necessary to make a diagnosis it would be difficult to resolve an understanding from the work of these clinicians.

Clinician Intra-rater Agreement

Clinicians difficulty in producing acceptable κ statistics in inter-rater testing extends to intra-rater testing as well. In most cases, intra-rater agreement addresses a clinician's ability to reproduce annotations on data they've previously seen. Ger-

Table 2.8: 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

ber et al., Grant et al., and Westhall et al. ran specific intra-rater experiments to contextualize the inter-rater results.

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 ask asked to follow the same classification scheme as the original study. The resultant κ statistics, Table 2.9, show the difficulty in a clinician agreeing with themselves. Compared against inter-rater agreement, Table 2.5, the intra-rater agreement is 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) is better than that of the inter-rater agreement (0.29 to 0.62). However, the intra-rater results suggest clinician

Table 2.9: 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

A is the worst performer. This in conflict with clinician A's inter-rater agreements, Table 2.6. The worst inter-rater agreements do not involve clinician A, but rather clinicians B, D, and F. These results suggest inter- and intra-rater agreement scores are poor tools for understanding a clinician's annotation ability.

Table 2.10: 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 2.11, scoring higher than inter-rater agreement, Table 2.7, is also exhibit by the clinician's test by Westhall et al.. Repeating the original experimental protocol 6 months later produced intra-rater very high classification agreement, Table 2.11. However, the κ statistic for highly malignant, 0.64, is lower than its inter-rater counterpart, 0.71. So despite each clinician improving their ability to identify the waveforms, they were unable to identify the

same waveforms as they did in the previous experiment. This speaks to nature of clinicians only in search of enough information to affirm a diagnosis or not.

Table 2.11: 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 2.8 represent less discrete facets of EEG waveforms. These features require qualitative analysis which increases the difficulty of classification consensus, exemplified by the abundance of slight and poor inter-rater κ statistics. Intra-rater agreement shows minimal improvement of κ statistics, while the averaged intra-rater agreement % is better than its counterpart. This suggests clinicians are capable of reproducing their work, but remain prevented from doing so by their innate biases thus limiting their κ statistics.

As a whole these studies indicate clinicians are consistent within themselves, and their cohorts, when classifying EEG recordings. Yet, that consistency does not appear to translate into producing gold standard datasets. While the results of each study offer suggestions as to why such consensus is difficult to reach, there is 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 impact the resultant κ statistics produced by clinicians. Their inability to come to agreement does not diminish their ability to diagnosis, but does limit the quality of data available to train algorithms.

2.2.2 Algorithm Classification

Clinicians utilize filtering and montages to enhance their ability to interpret EEGs and ML algorithms are no different. For both the clinician and the algorithm it

is necessary to select only the range of frequencies relevant to the brain. In that approach nothing changes, but algorithms require this filtered data to be converted into features. A clinician would apply a montage to the raw data producing features via differential electrode pairings. While the features for algorithms is often more involved.

Each study may utilize a unique feature set [14], borrow from existing work [94], or take the data in its raw form [95]. The task of feature development is beyond the scope of this work, but it is important to understand a few concepts related to EEG features. First, the term *epoch* is defined as the area in which features exist. Raw data is turned into features with epochs of n number of seconds. This forces trade offs between categorizing phenomena occurring rapidly, PDs, or slowly, such as sleep state. However, it enables data reduction by condensing all the samples in an epoch into a feature vector. Given the number of channels in a recording, their duration, and sample rate EEG recordings produce significant amounts of data. Dimensionality reduction is the other critical role of EEG features as it allows computational solutions to approach near real-time classification.

Thus features must excel at minimizing the amount of necessary data while ensuring nothing of significance is ignored. A difficult task which often sees features sets developed for specific use cases like seizures [13], BCIs [96], sleep [24], and other behaviors of interest such as alcoholism [76] and ADHD [97]. The combinations of features and epochs allows each study to focus on their specific goals, but makes a universal feature set difficult to define.

With each advancement in ML, the EEG community works to adopt the latest technique to their medium. As each technique matures it becomes another option for classifying EEGs. This is the goal of using I-Vectors and follows in the path of techniques like K-Nearest Neighbors (KNNs), Support Vector Machines (SVMs), Neural Networks (NNs), and GMMs. Often a given a combination of features and data

may perform better or worse depending on the algorithm. Therefore the performance of these older algorithms is used to benchmark not only new algorithms, but also feature.

The following sections reviews algorithms that use *statistical models*, *supervised algorithms*, and *unsupervised algorithms*. Statistical models form the basis of numerous ML techniques and are frequently used to filter out artifacts via thresholding, detect ERPs, or interpret common spatial patterns (CSPs). Supervised algorithms use labeled data and *a prior* knowledge to build classifiers. Meanwhile, unsupervised algorithms leverage statistical modeling of large sets of unlabeled data to build classifiers. All of these approaches can be applied to data generated from sleep, seizures, ADHD, or BCI EEGs. While I-Vectors will not be tested in all of these areas, the increasing complexity of detecting artifacts, sleep, and motor control signals presents a complete picture of EEGs.

Statistical Algorithms

Statistical modeling of known EEG phenomena provides an easy platform for algorithm based classification. The type of modeling depends on the waveform, but all the classification follows the same binary labeling of inside our outside the model. These approaches are mathematically straightforward and require minimal data having been built entirely on the knowledge of clinicians. This makes them susceptible to variations in data and thus unreliable on data not accounted for in their initial modeling.

The use of an ERP known as P300 drives the most basic BCI platform of P300-spellers. A P300-speller detects response to auditory or visual stimulus enabling a person to spell words with their brain [73]. This phenomena is ideal for statistical modeling as it only requires brief training on a given subject to tune feature weights for acceptable performance [98].

Guger et al. [98] showed 5 minutes of training were enough to elevate the majority of the subjects over 60% accuracy, Table 2.12. The training period asked the subjects to spell specific words and then used Linear Discriminate Analysis (LDA) to tune the weights of the 8 electrodes. 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 2.12

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 is highly effective at enabling communication without requiring large amounts of data or processing beyond LDA and frequency filtering. The main drawback is 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 is also very specific to ERPs which is not commonly associated with medical abnormalities. However, statistical models exist for clinical diagnosis, Alzheimer's Disease (AD) [99], ADHD [100], and seizures [101].

Seizure classification is a principle driver of EEG research with a focus on seizure prediction. Chu et al. [13] apply *attractor states*[8] to EEG data in an effort to improve seizure prediction and detection on two datasets, the Children's Hopsital of Boston Massachusettes Institute of Technology Scalp EEG Database (CHB) and adult

[8]Attractor states are stable states which the data trends towards given its natural behavior. The concept originated from the work of Scheffer et al.[101], but is beyond the scope of discussion in this work.

seizures from the Department of Neurosurgery of Seoul National University Hospital. The data is split into 20 second half-overlapping channel-independent epochs that are converted to frequency banded Fourier coefficient features. These features are evaluated against models of seizure and non-seizure states with a seizure prediction horizon of 30 seconds.

Predictions on the training data had an average sensitivity of 90.20% and 86.67% on the testing data. As sensitivity decreased, so too did the average false positives per hour from 0.476 on the training data to 0.367 on the testing data. The highest false positives per hour were 1.667 and sensitivity for multiple subjects was 0.0%. The results suggest a simple model can predict seizure onset, but it is not robust enough to maintain adequate performance across all 17 subjects.

Another principle focus of EEG research is sleep detection. Warby et al.[24] compared the performance of six statistical sleep spindle algorithms {a1[102], a2[103], a3[104], a4[105], a5[106], and a6[55]} to clinicians and non-experts. A sleep spindle dataset consisting of 32,112 25 second single channel epochs from 110 healthy subjects was curated to provide testing and verification data. A gold standard verification set was built from 2,000 epochs evaluated by an average of 5.3 clinicians.

Each of the algorithms applied different flavors of energy thresholding (Root Mean Squared (RMS), PSD, or Fast Fourier Transform (FFT)) on a bandwidth (9-16Hz) filtered portion of the epochs. The performance of the algorithms, Table 2.13, was not in agreement with the gold standard (GS), but the algorithms did agree with the automated group consensus (AGC). Overall, the algorithms were found to be the weakest and the clinicians the strongest at classifying sleep spindles. With non-experts performing better than the algorithms, the work suggests statistical based algorithms are not ideal for waveform classification.

In Huang et al. [107] it was found that classification of AD against control subjects was 84% correct using a combine alpha (8.0-11.5Hz) and theta (4.0-7.5Hz) global field

Table 2.13: 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

potential (GFP), a generalized EEG amplitude. The study evaluated 15 2 second epochs from 93 subjects to understand the variability in EEG patterns of subjects with AD, mild cognitive impairment (MCI), and healthy controls. They produced FFTs from each epoch and decomposed them based upon their GFP across frequency bands (delta (1-3.5Hz), theta, alpha, beta 1 (12-15.5Hz), and beta 2 (16-19.5Hz)). These features are localized based upon their location: antero-posterior (Loc-X), left-right (Loc-Y), and superior-inferior (Loc-Z). The results of the study are shown in Table 2.14.

In addition to AD it is possible to classify ADHD through a subject's theta beta ratio (TBR) [108]. Lenartowicz et al. [108] review multiple approaches for distinguishing ADHD patients from controls through the use of temporal and spatial features and ratios between energy in bands or channels. The studies found a range of performance when using TBR as a discrimination metric. Monastra et al. [100] reported an accuracy of 91% (90% sensitivity, 94% specificity) while Buyck et al. [109] 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

Table 2.14

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)

TBR a poor classification metric. Despite clear clinical utility in using EEG recordings for ADHD diagnosis [110], the condition was not understood well enough to rely solely on a statistical modeling.

Incidentally, Buyck et al. found that TBR did make an excellent, AUC 0.965, discriminator for age classification. This shows the difficulty in attempting to find an ideal feature set for classification as the same features can express multiple conditions. Detecting, and often correcting, artifacts is an exemplary example of this problem as artifacts often present with properties similar to waveforms of interest [8].

The most common artifacts (eye blink, muscle artifacts, and eye movements) are caused by the subject making them difficult to mitigate during recording. Jung et al. [111] indicate the overlap between artifacts and waveforms of interest prevents many novel artifact detection techniques from broad application. In their approach, the performance of independent component analysis (ICA) is compared against principal component analysis (PCA). Both of these algorithms are effective methods for

performing factor analysis which attempts to produce a simplified set of underlying statistical distributions.

From these decomposed distributions Jung et al. show can be effective, but ICA produces better results. Both techniques are predicated on the distributions having linear independence and non-Gaussian distributions. While these assumptions may be correct for artifacts, it is not assured for all types of EEGs waveforms. Delorme et al. [112] devised additional criteria[9] specific to artifacts to improve artifact classification building on Jung et al.'s work. They applied six thresholding schemes to raw data and data processed each with ICA.

Their results, Figure 2.5, showed that applying ICA improved artifact detection compared to the raw data, specifically on artifacts with larger signal to noise ratios. Despite the increased complexity of the algorithm, performance gains were not seen in all cases. This suggests that artifact type and detection scheme may dictate performance especially at lower signal to noise ratios.

The success of these ICA based approaches lead to 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) [113]. ADJUST and FASTER both provide a universal artifact detection and rejection technique applicable to multiple types of EEG data. They were developed in response to the time it took clinicians to review and clean EEG recordings for common artifacts such as blinks, eye movements, electromyography (EMG) artifacts, and white noise.

[9]They compared five methods to determine how best to identify artifacts within a recording. (1) Exterme 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.

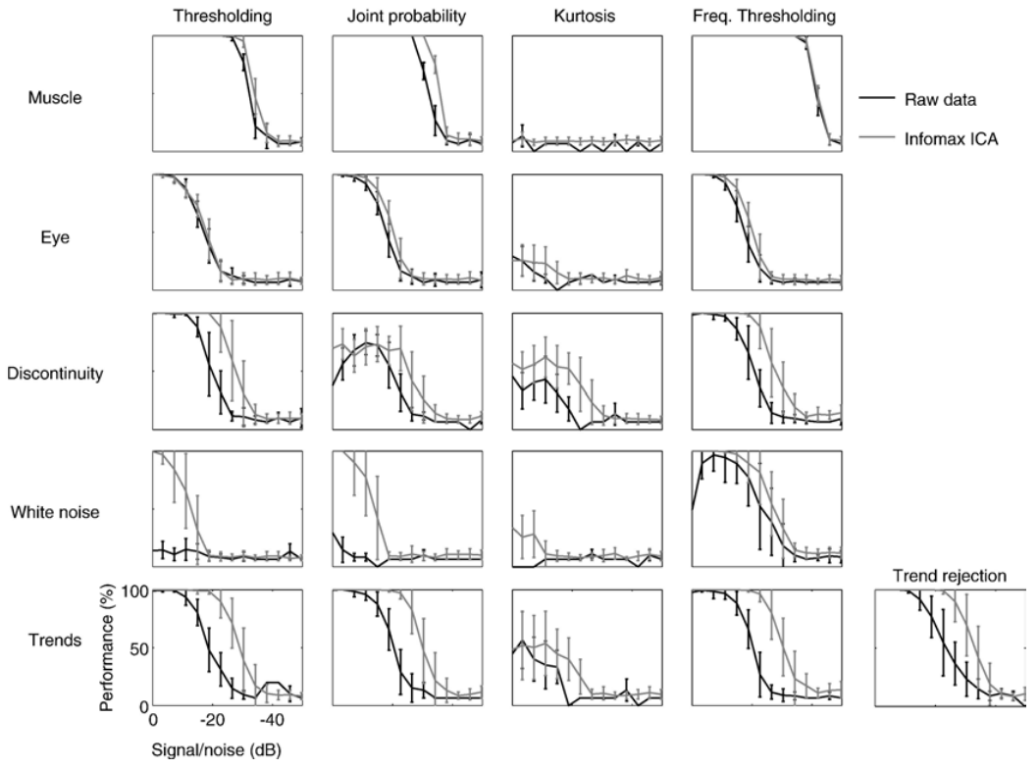

Figure 2.5: Classification performance of thresholding approaches based upon the signal to noise ratio of the artifact and the signal.

FASTER evaluates 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). The process is complex, but they achieve 60% sensitivity and 97% specificity on the epochs from 47 subjects recorded with 128 channels, Table 2.15.

Table 2.15: 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

[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.

ADJUST applies ICA on the filtered electrode data. It then uses spatial and temporal feature extraction to classify and remove artifacts. The process is less complex than FASTER, but is able to match a clinician's cleaned dataset with 95.2% accuracy. They show subject independent artifact defection is possible as the training dataset (21 subjects) and validation dataset (10 subjects) are comprised of unique subjects. This contrasts with the failure of ADHD TBR to discriminate over unique datasets.

Artifact detection and correction continues to be an active research topic, but the reliance on ICA remains. Mahajan et al.[38] report 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 results are promising, but come from one dataset focused on classifying one type of artifact. Classifying a single well defined waveform on a limited dataset appears to be the extent of these ad-hoc techniques. Attempts at classifying multiple artifacts across varied datasets with simple statistical modeling fails to provide robust classification. To develop algorithms capable of matching clinician performance researchers rely on ML algorithms.

Supervised Algorithms

Supervised ML algorithms build statistical models from labeled datasets. Instead of applying ICA and thresholding based on the independent components, supervised algorithms produce decision surfaces for each class of waveform. A decision surface allows the relationships between the waveform's features to dictate classification, instead of relying on a features handpicked by clinicians. This provides supervised algorithms with freedom in feature selection as the clinicians are in control of the classification label, but not how to make that classification.

For EEGs classification, classes are modeled from the epochs labeled by clinicians. The algorithms attempt to emulate the clinician's classification by making assumptions about the epochs and features of the targeted class. This exposes a limitation of supervised learning: The algorithms must be shown what to classify making their success dependent on the training data. As such supervised ML classification algorithms of well known phenomena (artifacts, seizure, and sleep) are more prevalent and robust than those from phenomena less well understood and less documented like (BCIs, emotions, and workload).

The classification of sleep relies on detecting waveforms unique to the stages of sleep: k-complexes and sleep spindles. There are also generalized behaviors concerning brain activity that accompany these specific waveforms [80]. Changes to the dominant EEG rhythms was previously shown to aid in ADHD and age discrimination. For example, dominant (>50%) alpha rhythms are indicative of wakefulness when classifying for sleep. Each stage of sleep contains a mixture of unique waveforms and shifts in the rhythms of the brain. Stage 1 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 make adaptation of supervised ML algorithms straightforward. In Schluter et al.[34] the stages of sleep are classified with Decision Trees (DTs) by bagging[11] on an array of physiological data[12]. The resultant classification of the 33,542 30 second epochs drawn from 15 subjects is shown in Table 2.16. Identifying wakefulness and REM sleep occurs with the highest accuracy, but the intermittent

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

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

stages of sleep are harder to classify. These results incorporate the use of data beyond EEG suggesting EEG may not be sufficient for accurate classification.

Table 2.16: 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

In Radha et al. [81] similar performance to Schluter is seen using only EEG data. This data comes as 34 features per 30 second epoch from 10 healthy subjects. Two supervised algorithms, Random Forest (RF) and SVM, classify the epochs into REM sleep and 3 stages of non-REM sleep (N1,N2,N3). The records are annotated by a trained clinician providing a reference that allows a κ statistic to be associated with each algorithm's performance, Table 2.17. Prior to classification the feature set was optimized for channel (F4-A1), epoch duration (30), and number of features (20).

Table 2.17: 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 2.16.

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	

These results confirm that performance is comparable to studies utilizing more data, like Schluter in Table 2.16. The moderate to substantial κ statistics show that the algorithms perform well, but it is likely that feature optimization driving performance. Since sleep is not a unique phenomena and represents a major changes in brain activity, the necessity of channel and feature optimization suggests the approaches are not robust given the depth of knowledge.

Similar to sleep, the behavior of seizures can be categorized into four stages: *normal* indicative of a normal healthy state, *pre-ictal* indicative of a build up to a seizure, *ictal* indicative of an active seizure [114], and *post-ictal* indicative of the time following a seizure [13]. Accurate detection of these, specifically pre-ictal, can help improve diagnosis and treatment of epilepsy [5].

Seizure classification is one of he most common tasks automated by algorithms [5]. There are commercial products available, but their performance is middling at best [115]. Efforts to improve classification focus mainly on developing better features beyond the common FFT based frequency band powers [14, 52, 13] and algorithms [114, 25, 116]. These efforts are predicated on, and thus limited by, the data available and the quality of the clinician's annotations.

Wulsin et al.[14] utilize raw data and a diverse feature set[13] associated with the signals to provide a basis for annotation. Part of the work sets out to identify the strongest features availableto improve performance on the channel-second epochs. However, the strongest classifications utilize the raw data over the various feature sets. In addition to the feature analysis, four classification algorithms (DTs, SVMs, KNNs, and Deep Belief Networks (DBNs)) are evaluated, with SVMs being the top performer, Figure 2.6.

[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

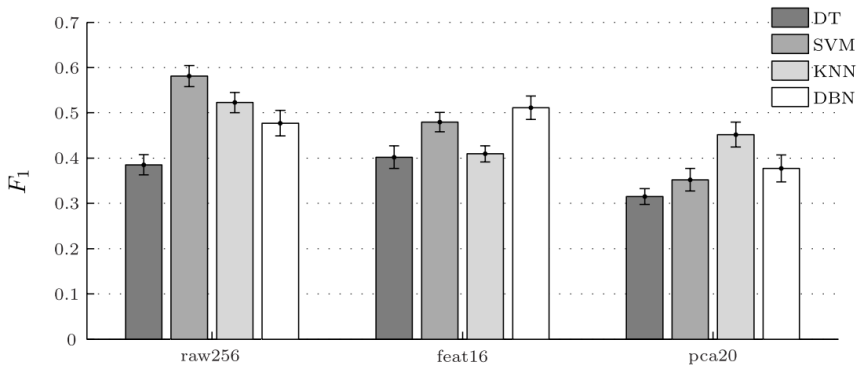

Figure 2.6: Wulsin et al. evaluate algorithm peformance based upon the F_1 measure, where $F_1 = 2 * (sensitivity * precision)/(sensitivity + precision)$. The results are presented to compare the algorithms and feature sets against each other. The feature sets are comprised of: *raw256* represents the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

Bajaj et al.[52] applied emperical mode decomposition (EMD)[14] features feeding a Least Squares Support Vector Machine (LS-SVM) classifier to identify seizures in 100 23.6 second channel based epochs from 5 subjects. EMD decomposes the nonlinear and non-stationary components into intrinsic mode functions (IMFs) making it ideal for generating features from non-stationary data like EEGs. The two dominant IMFs, amplitude modulation and frequency modulation, provided sensitivity and specificity of 100% with average sensitivity and specificity of 94%.

Acharya et al.[114] evaluated six supervised ML algorithms, Fuzzy Sugeno Classifier (FSC), SVM, KNN, Probabilistic Neural Network (PNN), DT, and Naive Bayes Classifier (NBC), and one unsupervised, GMM. Each algorithm was trained on four features derived from entropy calculations: Approximate Entropy (ApEn)[118], Sample Entropy (SampEn)[119], and S1 entropy and S2 entropy[120]. The data from 5 healthy subjects and 5 epilepsy subjects was pre-processed to produced 200 healthy, 200 pre-ictal, and 100 ictal artifact free single channel 23.6 second epochs.

[14]A detailed review of EMD is omitted, but if interested the work of Huang et al.[117] introduced technique and its applications.

Table 2.18: 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

Each algorithm's performance is shown in Table 2.18. Sensitivity and specificity are similar across the algorithms, but the best accuracy is achieved by the FSC classifier. The separability of the trained seizure states (healthy, pre-ictal, and ictal) produced a p-value less than 0.0001 for each of the features. With only 10 subjects in the study and strong discrimination from the features, it is hard to assess the strength of the individual algorithms given the lack of data diversity.

Differentiating the impact algorithms and features have on classification performance becomes harder as algorithms increase in complexity. This requires experiments deploying NN to benchmark themselves, as Ghosh-Dastidar et al. [25] did when testing a novel wavelet-chaos-neural network, Levenberg-Marquardt Backpropagation Neural Network (LMBPNN), on seizure datasets. The data was transformed into band specific features (standard deviation, correlation dimension, and largest Lyapunov exponent) for each of the 100 single channel recordings coming from healthy, pre-ictal, and ictal datasets. The 23.6 second 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).

Each frequency band was evaluated independently and then in various mixed-band configurations to find the optimal classification. The results provided an exhaustive analysis of the relationship between algorithm and feature set performance. The majority of the classification performance is similar so only the maximum accuracy is reported in Table 2.19. Despite the use of novel algorithms, the features appear to be the driving force of classification. LMBPNN has poor classification accuracy, below 50%, when using band-limited (0-60Hz) correlation dimension as a feature, but superior performance with mixed-band features.

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

Finding the right balance of features in terms of quality and quantity is a hurdle for NN EEG classifiers. Subasi et al. [116] used a small subject set (5 subjects and 4 channel per subject) that produced 500 5 second epochs for seizure classification. The epochs were labeled by two neurologists to produce a gold standard for evaluation of epileptic or normal waveforms. Epoch classification was carried out by multilayer perceptron neural network (MLPNN) with back-propagation, LMBPNN, and logistic regression on features produced by discrete wavelet transform (DWT). The features produced by the DWT span 6 frequency bands (0-3.125Hz, 3.125-6.25Hz, 6.25-12.5Hz, 12.5-25Hz, 25-50Hz, and 50-100Hz) that loosely align with standard EEG frequency bands.

Table 2.20: Performance comparison of Linear Regression versus

Classifier	Accuracy (%)	Specificity (%)	Sensitivity (%)
Logistic Regression	89	90.3	89.2
MLPNN w/ backprop	92	91.4	91.6
LMBPNN	93	92.3	92.8

The performance of the three algorithms is presented in Table 2.20 with LMBPNN coming out on top. These results come from one small unique dataset of ten subjects which leaves doubt about performance on more generalized populations. The common performance of the three algorithms suggests that feature selection may be more important than algorithm selection. All of these incongruities reduce the likelihood of educational advancement in terms algorithms, features, and understanding of EEGs from this work.

Unsupervised Algorithms

Unsupervised ML algorithms operate in a similar fashion to supervised ML algorithms, except they do not require labeled data. This means unsupervised algorithms develop classification schemes from the data without any external knowledge. The decision surfaces they create separate classes found by the algorithm as it maps the statistical properties of the data. A downside to this approach is that it requires a large amount of data to build representative models. The distribution of the data is important as well because the lack of labels means underrepresented events may not be included in the models despite being significant.

Given the need for larger datasets, the use of unsupervised classification of EEG recordings is less frequent than supervised classification techniques. Often an unsupervised algorithm is used as a comparison point, Acharya et al.[114] showed GMM produced competitive accuracy and sensitivity, but not specificity, Table 2.18. However, Ghosh-Dastidar et al. [25] used k-means clustering and found it performed

worse than the other algorithms, Table 2.19. As unsupervised techniques are more dependent on the dataset than supervised techniques, it is not uncommon to see large variance in performance.

Gabor et al.[121] 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 on features produced by a wavelet transform using 4 second epochs built from the 10 channels of each recording. A separate feature set using 8 second epochs was used, but the duration was too long and masked short seizures.

In total, 62 seizures were captured from the 24 recordings of which the algorithm detected 56 (90%). However, the average false positives per hour (0.71) produces 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. In this case the age range (<1 to 43 years old), small training set (5 of the 24 recordings), and epoch duration could be impacting performance.

Not all unsupervised algorithms focus on classifying the data, as some are deployed for dimensionality reduction. The most common occurrence of this is the use of unsupervised LDA in areas where clinicians' skills are weaker such as BCI [37]. LDA is factor analysis, in the family of ICA and PCA, but its behavior and application finds it labeled as an unsupervised classification technique.

Vidaurre et al. [37] uses three flavors of LDA to enhance the performance between BCI training and feedback sessions. The core discrimination technique, $LDAI$, is given as changes in the pooled mean (PMean) between the features seen in the training and feedback data. Expanding on this, $LDAII$ incorporates updates to the covariance matrix with PMean and $LDAIII$ scales the mean and covariance CSPs.

[15]A detailed review of SOMs is omitted, but if interested the work of Kohonen[122] 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.

These techniques were compared against a supervised LDA and mean to determine the feasibility of the unsupervised techniques.

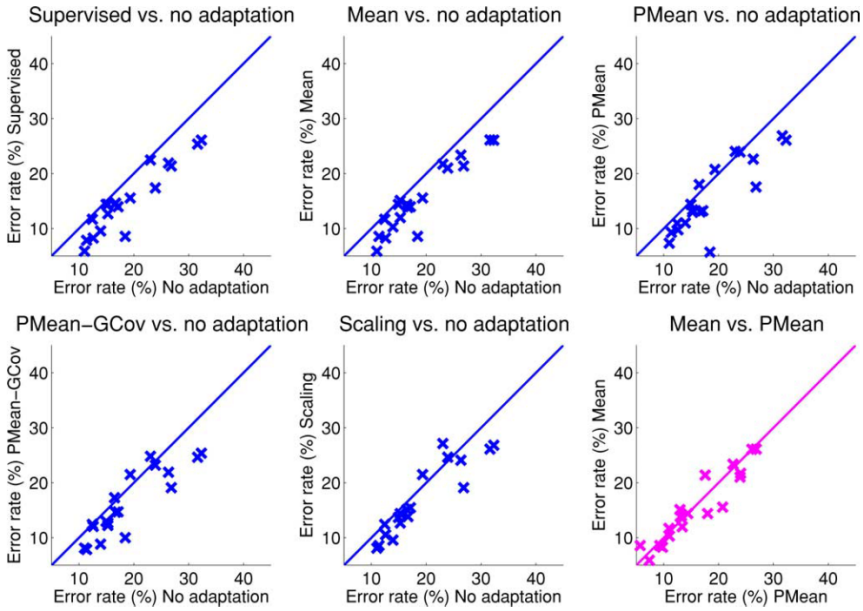

Figure 2.7: The comparative error rates between the supervised and unsupervised adaptation techniques through changes in the error rate. The pink plot shows the difference between a labeled, mean, and unlabeled, PMean, classification.

The unsupervised techniques were proposed as a simpler alternative to the work of supervised techniques. Their evaluation therefore depends on the being on par with the supervised techniques which is shown to be the case in Figure 2.7. The data for the comparison comes from calibration of 19 recordings of 10 subjects and shows the supervised algorithms slightly out performing the unsupervised algorithms.

Further testing, with a dataset of 80 recordings from 80 subjects, shows that PMean based algorithms meet or exceed the performance of the state of the art supervised approaches during feedback. The unsupervised technique exhibits robustness as one class is removed from BCI feedback and outperforms the supervised algorithm in Figure 2.8. These results are important because clinicians seldom label BCI datasets and BCI recording sessions are more dynamic than seizure or sleep recording sessions.

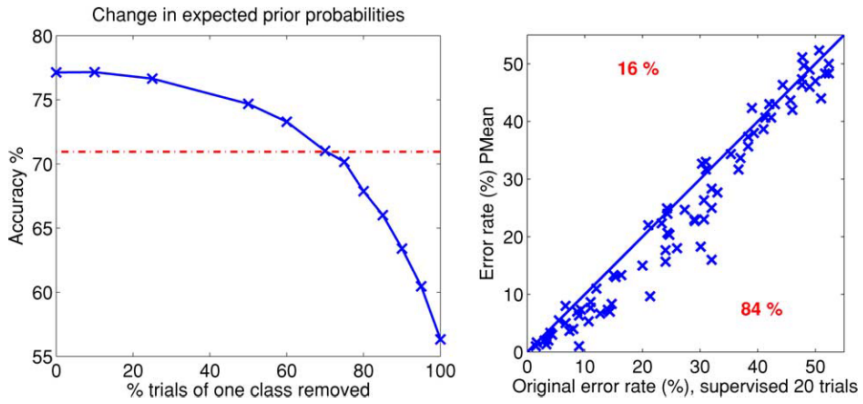

Figure 2.8: Performance on feedback data after training for supervised adaptation and unsupervised PMean adaptation. The (left) impact of removing one class from the feedback dataset for the supervised algorithm (red line) and unsupervised algorithm (blue line). The (right) error rate between the two algorithms during the online feedback experiments.

2.2.3 Bio-metric Applications

The use of EEG recordings as a means of bio-metric identification is not a new, but has only recently gained momentum. Initial attempts were able to discriminate EEG behavior between individuals and between different brain conditions [123]. This work did not have discrete waveforms to find or frequency ratios to calculate, but instead relied on direct comparison between subjects. Stassen [124] 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, remains in use today [57].

Advancement of EEGs as a bio-metric tool focuses on the statistical properties of each subject and is not reliant on clinician input. The independence from clinicians brings a need to control the dimensonality of the data which requires finding novel ways to enhance the distinctions between subjects. This makes bio-metric applications open-ended as they cannot rely on the decision surfaces used for known disorders. This use case aligns with the intended application of I-Vectors for differ-

entiating epochs, channels, and subjects from each other based upon their inherent statistical properties.

The initial efforts of those like VanDis et al. and Stassen focused on subjects at rest with eyes closed and open. Similar experiments are still carried out in the the work of La Rocca et al.[31, 62, 63] in order to optimize the accuracy, channels, features, and speed of subject verification. Active state recordings have subjects perform mental tasks such as imagining performing hand movements [125, 66], imagining speaking syllables [64], or reading text[126].

Active and resting based data emphasizes that the qualities of subject authentication and identification exist regardless of brain state. Other works have gone so far as to suggest a genetic basis underlies this separability [65, 36]. While interesting, the genetics of brain uniqueness expands beyond the scope of this work. By focusing on the techniques and results of active and resting based data studies comparisons can be drawn between the structured waveform based annotations of artifacts, seizures, and sleep.

Resting Recordings

The work of La Rocca et al.[31, 62, 63] focus on developing a novel set spatial and temporal patterns to improve subject recognition accuracy. Brigham et al.[64] work on data with imagined activities to test applications of subject identification during mental tasks.

In [31] electrode sets of 2, 3, and 5 from 56 channels positioned in accordance with the 10-20 system. Autoregressive stoichastic modeling and polynomial regression are used to match each 3 second epoch across the 6 standard EEG bands. Performance varies as a function of electrode set and EEG band with the trend of increasing electrodes improving performance. Regardless of electrodes, the alpha band provides the strongest classification accuracy. Peak performance of 98% classification accuracy

is found with 5 channels using signals across the alpha, beta, delta, and gamma bands. The best single band performance across 5 channels is 83% using the alpha band.

The follow on work in [63] uses 'bump' modeling to reduce the amount of data from the 10-20 layout into a parametric model. These bumps are filters that enabling sparse encoding. By generating a vector that controls mapping/weights of the bumps the vectors act as the features. These vectors are classified with LDA based upon features generated from groups of three channels drawn from the six standard EEG bands. The training and testing sets are curated to provide overlapping frames, jointed, and without overlapping frames, disjointed. This distinction shows the impact of frame overlapping with the beta band performing best, 95% jointed and 74% disjointed and the alpha band a close second, 96% jointed and 67% disjointed. Not surprisingly performance is improved in the overlapping dataset.

In [62] the work is expanded beyond spatial patterns to temporal patterns. Instead of focusing on specific regions, they direct attention to changes in power spectral density over 1 second epochs. This forgoes their earlier attempts at reducing the amount of data to be analyzed and instead produces Gaussian mixture distributions as feature vectors. These vectors are evaluated via Mahalanobis Distance (MD) for classification. Using the results for each region of the brain, classification accuracy reached 100%.

Active Recordings

In Marcel et al.[125] a nine subject dataset is matched based upon their brain activity performing three mental tasks. These mental tasks require the subjects to imagine carrying out the following actions: moving their left hand, moving their right hand, and speaking words with a common leading letter. Epochs of half seconds with 50% overlap were turned into PSDs based upon their FFT. The resultant PSDs were spatial filtered over the 10-20 electrode configuration with a surface Laplacian

function. These features were trained on GMMs to produce baseline models for subject verification over a range of mixtures. Evaluation scores were reported as half total error rate (HTER) generated from the false acceptance rate (FAR) and false rejection rate (FRR).

$$HTER = \frac{FAR + FRR}{2} \tag{2.2-3}$$

The results, Table 2.21, of the left and right hand authentication of the subjects suggests performance is improved with an increasing number of Gaussian mixtures. These results represent the largest data set used, collected over three days. Results using smaller sets showed the imaging word task performed worse for authentication than the hand tasks.

Table 2.21: The FAR, FRR, and HTER of imagined hand tasks as a function of Gaussian mixtures.

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.[66] phase synchronization is used for identifying subjects. The PhysioNet resting eyes closed and resting eyes open trials were split into the standard EEG frequency bands and segmented into 12 second non-overlapping epochs. Finding the phase lag index (PLI) relationship between all the channels of an epoch produces distinct mappings between subjects. These topologies are reduced via Eigenvector Centrality (EC) to produce a feature vector for each epoch. The Euclidean Distance

(ED) between each feature vector informs decisions of similarity between the subjects for a given frequency band.

Table 2.22: 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.[64] explored subject identification using two unique data sets using the same approach. One data set consisted of Visually Evoked Potentials (VEPs) in alcoholic and non-alcoholic 120 subjects. The other datset had 6 subjects uttering two syllables, /ba/ and /ku/. Artifacts were removed from each set and processed into PSDs of their respective trial lengths, 1 second 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 reported 98% accuracy and KNN reported 93% accuracy both with a 95% confidence interval. The syllable dataset provided slightly higher accuracy measurements of 99% with SVM and 98% with NN both at a 95% confidence interval. The strong performance across both datasets suggests the approach works well on a fundamental level, but given the small subject size for syllable dataset further testing should be carried out.

In Gui et al.[126] a more contemporary ML technique, Artifical Neural Network (ANN) using feed-forward, back-propagation, and multiplayer perceptron, is 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 are 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 are 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.

2.3 Identity Vectors

I-Vectors are mathematical models designed to reduce the dimensionality of UBMs [127]. UBMs reduce a dataset of f-dimensional feature samples into C clusters of f-dimensional GMMs. I-Vectors can then be created by enrolling distinct samples into a modeling process involving the UBM and a TVM built from the enrollment samples. Finally, those I-Vectors are evaluated against each other and testing I-Vectors, built from testing samples and the TVM, by the l-dimensional distance between them. For example, this technique can 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 [128]. In contrast, I-Vectors collapse those three models into just one. The principal I-Vector equation is

$$M \approx m + Tw \tag{2.3-4}$$

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.

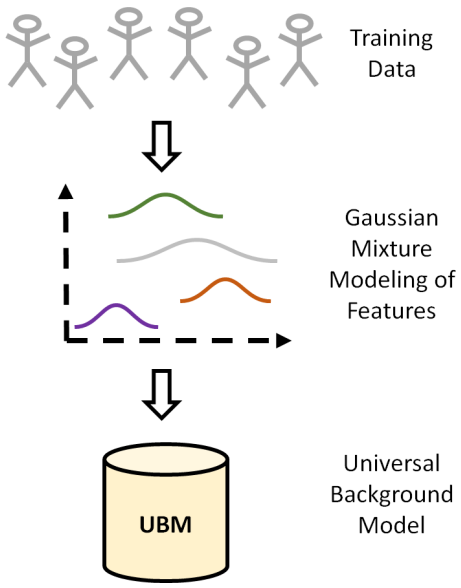

Figure 2.9: Training data is used to construct independent Gaussian mixtures for the f-features. The modeled feature space is separated into c clusters, each one a UBM. Taken as a whole these c UBMs provide a basis for the development of I-Vectors.

A typical application might involve determining whether EEG from a new patient suggests a diagnosis of epilepsy. First, a large randomized collection of training data would build a generic UBM, Figure 2.9. Then, sub-populations of enrollment data from known healthy and epileptic patients would be used to construct enrollment I-Vectors. Finally, the I-Vector from the new patient would be compared against the enrollment I-Vectors to determine which population it was more likely to match, Figure 2.10. 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.

A UBM models f-dimensional features by representing them with C independent Gaussian clusters [129]. In general, increasing the number of clusters captures more nuance, thereby potentially strengthening any ensuing discrimination. The UBMs provide dimensionality reduction by taking L epochs with f features each down to C mixtures of f features. As each feature has a mean m, variance σ, and weight ρ,

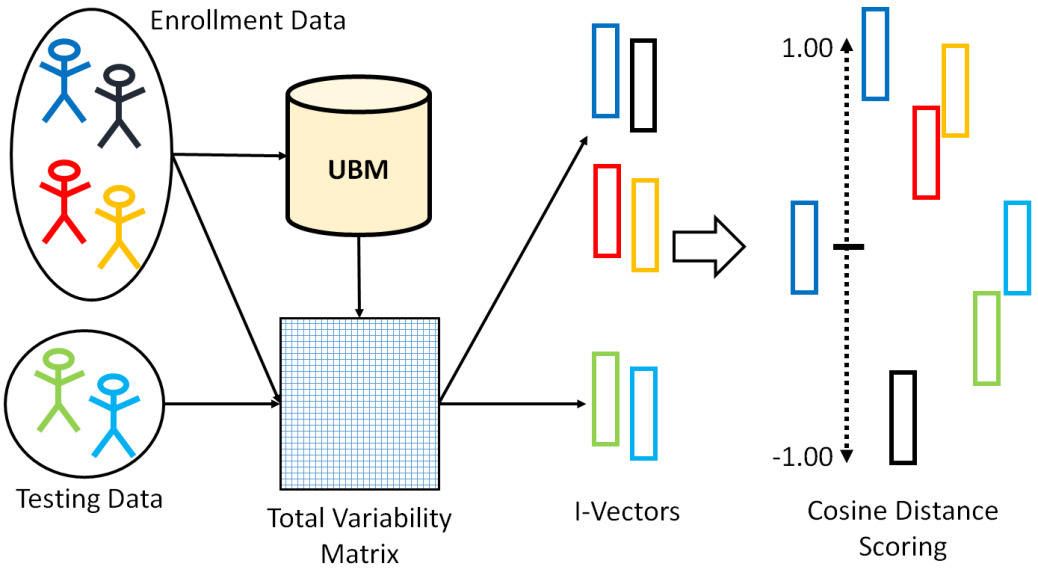

Figure 2.10: Using the UBMs as an initialization, the enrollment and training data are transformed into I-Vectors. This process is reliant on the creation of the total variability matrix randomly generated from the variances of the UBMs and refined by adaptation towards the means of the UBMs. The resultant I-Vectors are pairwise evaluated to find the CD between them to rank their similarity.

reduction benefits are seen when $L > 3C$. The UBMs can be characterized according to:

$$\Omega_{c=1...C} = \begin{cases} m(c) \\ \sigma(c) \\ \rho(c) \end{cases} \tag{2.3-5}$$

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.

These I-Vectors are of length $l = Cf$ with many residual elements that must be removed through the use of LDA. LDA creates a transformation matrix that removes dependent elements from the data which reduces the length of the I-Vector by one. This constrains the elements of the I-Vectors by driving them to a length of one less than the S targets in the enrollment data, thus $l = min(S - 1, Z)$ where Z is

on the order of 100s. The resultant I-Vector represents a location in l dimensional space. Within this space the similarity between two I-Vectors can be found via the CD between them.

2.3.1 Mathematics

The critical component of equation 2.3-4, 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 feedback from the produced I-Vector w.

$$T = \begin{bmatrix} T_1 \\ \vdots \\ T_C \end{bmatrix} = \begin{bmatrix} A_1^{-1} * K_1 \\ \vdots \\ A_C^{-1} * K_C \end{bmatrix} \tag{2.3-6}$$

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.

$$A_c = \sum_{s=1}^{S} N_s(t) w^{-1}(t) \tag{2.3-7}$$

$$K_c = \sum_{s=1}^{S} \hat{F}_c(s) * \left(w^{-1}(s) * T^H * \Sigma^{-1} * \hat{F}_c(s) \right)^H \tag{2.3-8}$$

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.

$$w(s) = \left(I + T^t \Sigma^{-1} \hat{N}(s) T \right)^{-1} T^t \Sigma^{-1} \hat{F}(s) \tag{2.3-9}$$

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 clusters resulting in $C \times f$ matrices. Each epoch, e, from the full epoch set L is evaluated to generate initial probabilities based on Ω for N and F.

$$\hat{N}(s) = \begin{bmatrix} N_1(s) & & \\ & \ddots & \\ & & N_C(s) \end{bmatrix} \tag{2.3-10}$$

$$\hat{F}(s) = \begin{bmatrix} \widetilde{F}_1(s) \\ \vdots \\ \widetilde{F}_C(s) \end{bmatrix} \tag{2.3-11}$$

$$\widetilde{F}_c(s) = F_c(s) - N_c(s)m_c \tag{2.3-12}$$

$$N_c(s) = \sum_{t=1}^{L} P(c \mid e_t, \Omega) \tag{2.3-13}$$

$$F_c(s) = \sum_{t=1}^{L} P(c \mid e_t, \Omega)e_t \tag{2.3-14}$$

This process resolves a suitable T after approximately twenty iterations of equations 2.3-6 to 2.3-9. Notice that equations 2.3-10 to 2.3-14 are needed only once to generate T. Creating I-Vectors from the enrollment and testing data follows equation 2.3-4 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.

$$w = (M - m)T^{-1} \tag{2.3-15}$$

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 [130]. These final I-Vectors can be compared pairwise using CD to determine similarity between enrollment targets and testing queries.

$$cos(\Theta_{w_1,w_2}) = \frac{w_1^t w_2}{\|w_1\| * \|w_2\|} \tag{2.3-16}$$

2.3.2 Success in Speech

The technique itself is well developed from its use on speech processing problems. All of this research outlines best practices for working specifically with speech signals so the main concern is how to adapt to a different data source. Fundamentally, evaluations of ML algorithms rely on tracking the sensitivity and specificity of each experiment and I-Vectors are no different. They perform inline with other approaches achieving over 90% sensitivity and 90% specificity [131].

The deployment of I-Vectors as a tool for speaker recognition/verification[132], language detection[133], accent detection[134], and speaker age[135] shows the trust the speech community has in the algorithm. I-Vectors were developed in 2011 at the Centre de Recherche d'Informatique de Montreal (CRIM) by Dehak, Kenny et al[130]. Prior to this work the group at CRIM developed JFA for use with speech data to address speaker and session variability[136]. I-Vectors come about as a natural extension from JFA which itself borrowed from previous research in mathematics.

One problem in adapting this work is that speech can easily discern when someone is talking and thus producing valid data. This cannot be easily replicated with 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 the work see in 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.

2.3.3 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)[137].

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

$$p(X|\lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\lambda) \tag{2.3-17}$$

this equation assumes that each component of the distribution is independent [16] 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.

[16] This often turns out to be untrue, but is a necessary assumption to provide a functional solution.

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

$$\bar{w_i} = \frac{1}{T}\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda) \tag{2.3-18}$$

$$\bar{\boldsymbol{\mu}_i} = \frac{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t}{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)} \tag{2.3-19}$$

$$\bar{\boldsymbol{\sigma}_i^2} = \frac{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t^2}{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)} - \bar{\mu_i^2} \tag{2.3-20}$$

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

$$\Pr(i|\boldsymbol{w}_t, \lambda) = \frac{w_i g(\boldsymbol{x}_t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{k=1}^{M} w_k g(\boldsymbol{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \tag{2.3-21}$$

2.3.4 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 [129]. 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 [138]. This allows a likelihood

calculation equation given a D-dimensional sample x to compare against the model,

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} w_i g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2.3-22}$$

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,

$$g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{\exp\Big\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\Big\}}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \tag{2.3-23}$$

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.

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 maximum a priori (MAP) estimation[138]. 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.

$$n_i = \sum_{t=1}^{T} \Pr(i|\boldsymbol{s}_t, \lambda_{\text{prior}}) \tag{2.3-24}$$

$$E_i(\boldsymbol{s}) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|\boldsymbol{s}_t, \lambda_{\text{prior}})\boldsymbol{s}_t \tag{2.3-25}$$

$$E_i(\boldsymbol{s}^2) = \frac{1}{n_i}\sum_{t=1}^{T} \Pr(i|\boldsymbol{s}_t, \lambda_{\text{prior}})\boldsymbol{s}_t^2 \tag{2.3-26}$$

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.

$$\hat{w}_i = \left[\frac{\alpha_i^w n_i}{T} + (1-\alpha_i^w)w_i\right]\gamma \tag{2.3-27}$$

$$\hat{\boldsymbol{\mu}}_i = \alpha_i^m E_i(\boldsymbol{s} + (1-\alpha_i^m)\boldsymbol{\mu}_i \tag{2.3-28}$$

$$\hat{\boldsymbol{\sigma}}_i^2 = \alpha_i^v E_i(\boldsymbol{x}^2) + (1-\alpha_i^v)(\boldsymbol{\sigma}_i^2 + \mu_i^2) - \hat{\boldsymbol{\mu}}_i^2 \tag{2.3-29}$$

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

$$\alpha_i^{w,m,v} = \frac{n_i}{n_i + r^\rho} \tag{2.3-30}$$

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 2.11. The models themselves can be compared against each other to determine similarity in addition to evaluating them against new data samples.

2.3.5 Joint Factor Analysis

JFA is a specific application of factor analysis[17] built by the speech community to separate an utterance M into information related to the speaker V, the communication channel U, and residual noise components D. The vectors y, x, z provide controlling weights to the matrices V, U, D[139].

$$M = m + Vy + Ux + Dz \tag{2.3-31}$$

[17]Factor analysis represents a field of math capable of separating signals into sub-components. Perhaps the most familiar tools in this area are ICA and PCA.

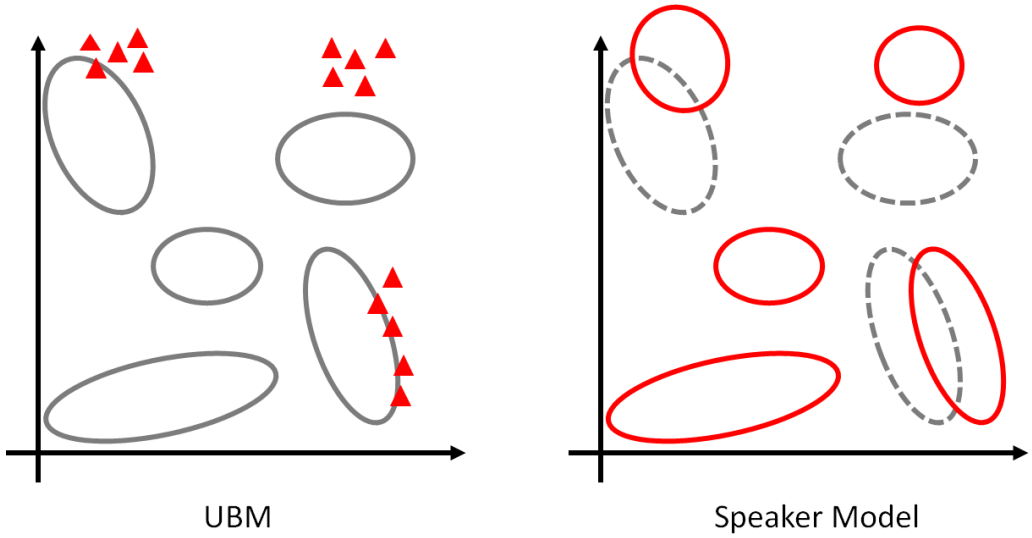

Figure 2.11: Results of MAP estimation when speaker data, red triangles, is applied to a UBM, gray mixtures.

The common model components, m, are a UBM built on the training data. Typically this is referred to as the speaker- and channel-independent supervector as the modeling process targets commonalities across the training data. If the training data is curated with intent to include only one channel type, all landline recordings instead of landline and mobile phone, then channel-independence is not assured. Additionally, if the recording sessions take place over a period of time then the UBM begins to account for session-independent features.

The sizes of the three matrices $\{V, U, D\}$ are dependent only on the space of the data. The rows of each matrix must be equal and contain enough variation to encapsulate all of the speakers' data. Naturally the rows should not meet or exceed the aggregate amount of data used during training. A one-to-one system negates all of the advantages offered by JFA, there would be no dimension reduction or factor separation.

Just as the number of rows reflects the variation in the training data, the number of columns reflects an assumption about the number of voices, channels, and residual effects in the recordings. The eigenvoice matrix V represents all speaker-dependent

components. If there are too many or too few rows in the matrix it will poorly map the voice space leading to decreased performance. This applies to the eigenchannel and residual matrices as well. As the size of these matrices impact performance it becomes necessary to optimize the rows of each matrix in some manner. Various metrics exist [140], but the nature of this process is beyond the scope of this work.

Solving for the eigenvoice, eigenchannel, and residual components matrices requires an iterative approach under an assumption from most to least influential. The process is mathematically the same as solving for the TVM. The difference is that when solving for V it is assumed that U and D are zero. As each matrix is determined, it is incorporated into equation 2.3-13. Therefore solving for V is the same as solving for T, but the solutions for U and D must adjust for the previously calculated matrices.

When solving for U the BW statistics must now take into account the channel *chan* and the subject s. This adds another dimension to the matrices and is dependent on separating the data by subject and channel. Thus L must be divided into time samples specific to subject-channels.

$$N_c(chan, s) = \sum_{t=1}^{L \in (chan,s)} P(c \mid e_t, \Omega) \tag{2.3-32}$$

$$F_c(chan, s) = \sum_{t=1}^{L \in (chan,s)} P(c \mid e_t, \Omega) e_t \tag{2.3-33}$$

Then F_c must account for the factors within V.

$$\hat{m} = m + V * y(s)$$

$$\widetilde{F}_c(chan, s) = F_c(chan, s) - \hat{m} * N_c(chan, s) \tag{2.3-34}$$

The updated N_c and F_c can be inserted into equations 2.3-10 and 2.3-11 respectively. Which allows the U and x to be substituted for V and y in the remaining equations.

This is then repeated a final time to address what is left once speaker and channel factors have been removed.

$$\hat{U} = U * x(chan, s)$$
$$\widetilde{F}_c(s) = F_c(s) - \hat{m} * N_c(s) - \sum_{chan \in s} \hat{U}(chan, s) * N_c(chan, s) \tag{2.3-35}$$

Again from this point the solutions for D and z are found in the same manner as before and substituted for V and y. With all the factors determined, the score for given target (tar) and test subjects can be evaluated.

$$\begin{aligned} \text{score} = &\big(V * y(tar) + D * z(tar)\big)^H * \Sigma^{-1} * \\ &\big(\hat{F}(test) - \hat{N}(test) * m - \hat{N}(test) * U * x(tst)\big) \end{aligned} \tag{2.3-36}$$

This score represents a linear distance between the target and test subjects. Depending on how the data is segmented the subjects could represent distance between words, phrases, or complete sentences of speech. There are many ways to control the discrimination process based upon the structure of the data and how channel factors are defined.

2.3.6 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[130]. 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, is 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 2.3-37. These parameters are dependent on each other and the training data. The speech community uses a definitive feature set [141], Mel Frequency Cepstral Coefficients (MFCCs), which evolved over time to become the gold standard [142]. This makes determining the number of features straightforward. Settling on an acceptable number of mixtures for the GMM is more difficult given the trade-offs between classification and computational performance[143, 144].

In many studies the number of mixtures is on the order of a base 2 number, often being set to at least 2048 mixtures[145, 131]. The optimization for the number of mixtures is 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.

$$\begin{bmatrix} M_1 \\ \vdots \\ M_f \end{bmatrix} = \begin{bmatrix} m_1 \\ \vdots \\ m_f \end{bmatrix} + \begin{bmatrix} T_{1,1} & \dots & T_{1,C} \\ \vdots & \ddots & \vdots \\ T_{f,1} & \dots & T_{f,C} \end{bmatrix} * \begin{bmatrix} w_1 \\ \vdots \\ w_C \end{bmatrix} \tag{2.3-37}$$

Critically, the TVM is not implemented to mimic utterances, but to map them instead. The technique allows I-Vectors to be the weights controlling the inclusion of a column of features. In this manner it is 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 understanding 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 parameter-

ization 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.

2.4 Machine Learning Algorithms

The breadth of potential algorithms, supervised and unsupervised, is too great to review in this context. Instead, a review of algorithms referenced in comparative works as well as those critical to the validation of I-Vectors will be reviewed. While all the algorithms can operate on the same features, their outputs tend to be unique. Their ability to classify, and in some cases cluster, and their training protocols are two defining characteristics of each algorithm.

For completeness a brief discussion of factor analysis (FA) is also included given the frequent use of LDA in supervised and unsupervised techniques. This also provides space to discuss PCA and ICA which will be used in deconstructing the relationships between I-Vectors, TVMs, and EEGs epochs.

2.4.1 Factor Analysis

At a base level I-Vectors reduce 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 [137]. While this is 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.

Principal Component Analysis

PCA finds the dominant components in a set of data by maximizing the variance of the given features [146]. 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.

$$\boldsymbol{V} = \frac{\boldsymbol{X}^T \boldsymbol{X}}{n} \tag{2.4-38}$$

$$\sigma_w^2 = \boldsymbol{w}^T \boldsymbol{V} \boldsymbol{w} \tag{2.4-39}$$

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.

Independent Component Analysis

ICA separates individual signals from collected by multiple receivers, commonly known as BSS [137]. The given example would be 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 $\boldsymbol{X}$.

$$\boldsymbol{X} = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} \boldsymbol{A} = \begin{bmatrix} a_1 1 \dots a_1 n \\ \vdots \ddots \vdots \\ a_n 1 \dots a_n n \end{bmatrix} \boldsymbol{s} = \begin{bmatrix} s_1(t) \\ \vdots \\ s_n(t) \end{bmatrix} \tag{2.4-40}$$

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} \tag{2.4-41}$$

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.

$$\boldsymbol{z} = \boldsymbol{V}\boldsymbol{x} \tag{2.4-42}$$

$$\boldsymbol{V} = \boldsymbol{E}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{E}^{T} \tag{2.4-43}$$

$$\boldsymbol{z} = \boldsymbol{V}\boldsymbol{A}\boldsymbol{s} = \hat{\boldsymbol{A}}\boldsymbol{s} \tag{2.4-44}$$

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.

$$\begin{gathered}\mathrm{kurtosis}(y) = \mathrm{E}\{y^4\} - 3(\mathrm{E}\{y^2\})^2 \\ \mathrm{kurtosis}(s_1 + s_2) = \mathrm{kurtosis}(s_1) + \mathrm{kurtosis}(s_2) \\ \mathrm{kurtosis}(\alpha s_1) = \alpha^4 \mathrm{kurtosis}(s_1)\end{gathered} \tag{2.4-45}$$

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.

$$|\text{kurtosis}(y)| = |q_2^4\text{kurtosis}(s_1) + q_2^4\text{kurtosis}(s_2)| \tag{2.4-46}$$

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.

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.

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

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

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

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

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}} \tag{2.4-47}$$

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.

$$S_X = (n - K)^{-1} \boldsymbol{S}_W \tag{2.4-48}$$

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.

Gaussian Classifiers

Once created, GMMs can be used as the basis for discrimination. As discussed in section 2.3.3, 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.

$$likelihood(x, \mu, \Sigma) = \frac{e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}}{\sqrt{|\Sigma|(2\pi)^d}} \tag{2.4-49}$$

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.

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

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)} \tag{2.4-50}$$

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

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B|A_i)\mathrm{P}(A_i)}{\sum_i^T \mathrm{P}(B|A_i)\mathrm{P}(A_i)} \tag{2.4-51}$$

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 $(0\% - 100\%)$.

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 [147].

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.

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 [148]. 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.

$$F_{sq}(x) = \text{sign}\Big[\frac{1}{2}(x-m_1)^T\Sigma_1^{-1}(x-m_1) - \frac{1}{2}(x-m_2)^T\Sigma_2^{-1}(x-m_2) + \ln\frac{|\Sigma 2|}{|\Sigma_1|}\Big] \quad \text{(2.4-52)}$$

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.

Hidden Markov Models

Unlike the previous classifiers, Hidden Markov Models (HMMs) take into account the temporal aspects of the data [149]. Just as KNNs and SVMs operate directly on the feature vectors so too do HMMs. However, they do so with regard to their previous classification state making them sensitive to temporal features. Their organization is similar to a multi-layer finite-state machine (FSM) where states feedback on one another. An *input layer* reads feature data from a sequence of samples which drives a

variable amount of *hidden states* in the *hidden layer*. The hidden layer then produces a classification via the *output layer*. Both the input states and output states are defined by the training data and its labels, but the hidden states can be chosen freely.

Training these models is usually a supervised process. A matching labeled output sequence is required for each training input sequence. Once trained, the model can be used to determine how likely a given input or output sequence conforms to the training data. For speech this would be taking a spoken sequence and resolving the utterances into phonemes or, in reverse, a sequence of phonemes could be used to generate utterances. In both cases the HMM is attempting to resolve the path through its network with the highest likelihood.

This approach can be adapted for unsupervised learning, but the result will be clustering and not classification. Without the presence of labeled output, the states of the HMM must be assumed. These states become the classes the data will separate into for classification and their properties must be estimated. Optimization of the network, via EM or similar, enables the estimated class properties to be refined to produce a HMM most likely to produce the data sequences.

Dirichlet Process

A Dircihlet Process (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[18] of the data Θ[150].

$$G \approx \mathrm{DP}(\alpha, G_0) \tag{2.4-53}$$

$$\Big[G(T_1), ..., G(T_K)\Big] \approx \mathrm{Dir}(\alpha G_0(T_1), ..., \alpha G_0(T_K)) \tag{2.4-54}$$

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 [151].

The version used in Wulsin et al.[49], Heirarchical Dirichlet Process (HDP), allows distributions to be drawn across multiple levels of the data at once. This exemplifies 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.

[18] 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.

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, 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 [152]. By training on ranked input vectors, in the case of Mueller et al. [152] 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.

CHAPTER 3

Methods

Those who fail to plan, plan to fail.

Attributed to Benjamin Franklin

Presently there is no active research being conducted on the integration of EEGs and I-Vectors. This means there is minimal guidance on how to adapt I-Vectors for use on EEG data. While the ultimate goal is subject and condition discrimination, optimization of the I-Vectors process must be carried out first. To adapt I-Vectors to work with EEGs three types of experiments will be performed: *Core Experiments*, *Principal Experiments*, and *Comparison Experiments*.

The Core Experiments will address the operational mechanics of I-Vectors with respect to what is an untested data type, EEG recordings. There is considerable variation among EEG recordings including the diversity of the subjects, variations in recording conditions, and characteristics of the recording itself. To mitigate the impact of these, the core experiments are carried out on identical training, enrollment, and testing datasets. This should mitigate issues with the data allowing to focus on the process of creating I-Vectors. The results will provide a baseline for developing an optimal set of parameters for producing I-Vectors from EEG data by sequentially testing each development parameter.

The Principal Experiments will validate the performance of the developed I-Vector process. Once internal best case parameters are found, the performance of the algorithm must be evaluated against alternative methods of classification. This serves to benchmark the technique, but also provides contrasting evaluations to identify

strengths and weakness in data classification. The weakness in unsupervised learning is that learning from its results are difficult unless there is a known link between the results and the data. Using other unsupervised algorithms provides necessary feedback for understanding where edge cases may exist in the datasets despite being unlabeled.

Finally, the Comparison Experiments will analyze how the I-Vectors are able to provide a robust feature vector capable of discrimination across multiple levels. The limits of similarity evaluations (subject characteristics and condition characteristics) will be reviewed based upon the results of the Core and Principal Experiments. With small tightly controlled datasets, in contrast to those of the Principal Experiments, it will be possible to track the influence of the raw data on the UBMs, TVMs and resultant I-Vectors. Through these experiments it should be possible to determine why I-Vectors offer improved classification and clustering performance relative to their counterparts.

3.1 Research Outline

The goal of this work is to develop I-Vectors as a suitable classification and clustering technique for EEG recordings. This requires (1) finding the optimal operating parameters for generating I-Vectors from EEG recordings, (2) comparing the technique to other commonly used EEG classifiers, and (3) explaining the comparative performance in terms of the strengths and weaknesses of I-Vectors as a modeling, classification, and clustering technique for EEG recordings. Each aim is driven by the use of a uniquely curated dataset. As the technique is an unsupervised ML algorithm, control over the datasets is the most direct way to influence the creation of I-Vectors. This is especially true after resolving the ideal operating parameters in accordance with RA 1. While RA 2 applies to multiple algorithms, the datasets are still the critical component driving the experiments.

3.1.1 Research Motivation

EEG recordings are rich with knowledge about their subject. Current classification techniques focus on the discrimination of specific components of the subject or the subject's condition. This mirrors the use cases of clinicians who diagnosis disorders and diseases based upon known EEG characteristics. These approaches are effective at treating patients, but leave a substantial amount of information undocumented in each EEG recording. Directing the diagnostic power of I-Vectors, as evidenced on speech data, to EEG data should enhance our ability to understand the human brain and thus the diagnostic skill of clinicians and algorithms.

3.1.2 Compositions of Datasets

Working with an unsupervised ML technique and unlabeled data requires robust datasets. Isolating specific components of the I-Vector generation process requires datasets constructed to address their specific functionality. UBMs treat the dataset as a singular entity so switching channels of one subject with another has no impact because the overall data content has not changed. Such a change would impact the subject I-Vectors as they are an aggregate of the channel data, but would not alter the individual channel I-Vectors. These relationships constrain, but also highlight, the way in which the dataset itself impacts the process.

A 100 subject subset of the TUH Corpus, consisting of 50 normal and 50 abnormal recordings, and the 109 subject PhysioNet dataset form the basis of the experimental data. This allows for three major datasets: the TUH Corpus dataset, the PhysioNet dataset, and a *Combine Dataset* built of the TUH Corpus and PhysioNet datasets. Each major dataset will be tested as a subject and channel classification dataset. In addition these six datasets will be further refined into a *Partitioned Dataset* that is organized by trials. The trials will slice the channel and subject datasets into fourths which is necessary for the LOOCV shown in Table 3.1.

3.2 RA 1: Optimal Operating Parameters

Determining the operating parameters requires a series of experiments iterating through the controllable parameters of I-Vector generation process. Each of these parameters will be swept one at a time over a range of values to find the optimal setting.

- the number of the UBM mixtures
- the number of EM iterations used to generate the TVM
- the allowable dimension of TVM rows
- data segmentation: full subject, full channel, and trials (partial subject and partial channel)
- feature influence

The impact of each parameter can be seen in different facets of the algorithm, so each one can be evaluated independently of the rest.

First, the significance of allowable UBM mixtures will be determined by producing UBMs of sizes {2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096} for each dataset. The process of generating a UBM is independent of how the data is partitioned, so each UBM must only be built once. Each mixture will be evaluated based upon the performance of its Gaussian Mixture Model-Universal Background Model (GMM-UBM) classifier through *verification testing*[1]. The quality of data modeling by each UBM will also be evaluated in terms of percentage of full dataset coveraged and percentage of overlapping mixtures.

The second set of experiments will assess the ability of each UBM to produce a sufficient TVM through EM. The implemented iteration process is suggested to have

[1]Verification testing is when the training and testing datasets are the same.

10 to 20 iterations, but that pertains to testing done on speech data. Therefore a more definitive iteration count is necessary for EEG data. The rate change of the matrix condition number along with the percent difference in mean, row variance and column variance of the TVM will be used to determine an acceptable minimum number of iterations. Development of the TVM is dependent on the BW statistics of the enrollment data creating an additional degree of freedom in addition to the UBMs.

Understanding the training process of the TVM then allows the third parameter, the number of rows in the TVM to be explored. Based upon the number of entries in the enrollment data the final number of TVM rows is limited to prevent a direct mapping between a given row and an enrolled target. This is achieved through LDA so the initial number of rows provides flexibility in the eventual transformation. The number of rows will range across {10, 50, 100, 200, 400, 800, and 1600}, but will not exceed the total $f * c$ for the chosen UBM where c is the number of mixtures and f is the number of features per mixture. I-Vector classification performance before and after LDA will be used to track the influence of number of TVM rows.

The influence of data organization and features must be measured through the classification performance of the I-Vectors. Assessing the performance of full and partitioned subject/channel dataset classification is necessary to provide performance benchmarks for feature influence and the comparison experiments. Once created, the UBM and TVM can be permuted to factor out specific features. In doing so it becomes possible to track the influence of each feature on classification and clustering performance. Using the optimal UBM and TVM combination will provide an understanding of impact of features for each dataset.

3.2.1 Core Experiments: Optimal Parameter Settings

The Core Experiments provide the basis for all future comparisons by evaluating thedatasets on the channel and subject level. For each experiment the training, enrollment, and testing does not change so all results can be attributed to the parameter under investigation. These experiments are estimated to take 5 weeks which is largely attributed to the run-time of the larger datasets' parameter sweeps.

The first experiments require each dataset to produce I-Vectors as the UBM mixture size is swept. There are ten datasets to evaluate: full subject, full channel, four partial subject and four partial channel. Each of these will produce equal error rates (EERs) and CDs for their I-Vectors. In addition, GMM-UBM evaluation will report back its own EER for each UBM providing a worst case comparison point for I-Vector performance. Determining the optimal mixture number will be based entirely on the reported EERs from the I-Vector evaluations. The best mixture counts for each dataset and in a generalized case will be recorded and used in subsequent experiments.

The second experiments focus on the development of the TVM through EM of the statistical characteristics of the datasets. Each of the optimal mixture sizes will undergo a prolonged EM to allow for the discovery of a local, or global, minimum error. This process will track percent differences of the row and column variances, the matrix condition number, the mean squared error (MSE), and the EER of the resultant I-Vectors for each iteration of EM.

The fourth experiment focuses on the number of allowable TVM columns. The number of rows in the TVM is a function of the number of features and UBM mixtures, but the number of columns is initially user controlled. Eventually the number of columns becomes the number of elements in the I-Vector making this process the a critical step in dimensonality reduction. The only constraint is that the number of

rows should not equal nor exceed the number of rows. Thus limiting the testable range of by the number of mixtures in the UBM.

Upon completion of these experiments, the final task is to assess the impact of the 26 chosen features. Through sequential removal of each feature from the dataset, UBM, and TVM new I-Vectors andEERs can be produced to track their influence. This should only need to be performed on the optimal UBM and TVM for each dataset. A review of the ordinal ranking of I-Vectors may be useful to assess clustering performance on the channel and subject level across each dataset.

3.2.2 Core Experiments: Justification

To determine an optimal setting for the number of mixtures produced by the GMM process, each training data set will be swept over a range of mixture sizes. A similar experiment was performed as part of the preliminary research, but it used one data set and only tested up to 1024 mixtures. With the Core Experiments additional datasets will be used and the range of mixtures will be increased to: {2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096}.

In essence, the mixtures are thought to be the features of data. If the number of underlying features is know prior to TVM creation then generating an adequate TVM would be simple. However when the underling components are not known, the process to create a TVM is unsupervised which introduces additional uncertainty forcing a wide net to be cast for the GMM creation process. To discriminate against S subjects, the thought would be that at least S mixtures must be produced. Early experiments showed this was partially correct depending on the search scope and the type of data being processed.

When reviewing the PhysioNet results, Figure 4.5, it was noted that the EER for subject-trial discrimination improved once the cluster size exceed the number of unique trials. So within a subject the trials were the strongest decision surface,

however the same response was not seen in the subject or trial independent results, Figure 4.4. A trend of increasing mixture size leading to improved performance is suggested, but there is no significant improvement when the number of mixtures exceeds the number of subjects or when the number of mixtures exceeds the number of trials.

From these results alone it is not clear which mixture size is best nor do they agree with the suggestion that mixture size is a proxy for underlying features. These are the reasons the experiments on mixture sizes are carried out again, but this time with a larger data and more varied dataset over an increased range of mixtures. To capture differences between the data sources, each subject will be run as an individual experiment along with a comprehensive experiment using all subject data.

3.3 RA 2: Comparative Algorithm Performance

The results of RA 1 provide the necessary framework to compare I-Vector classification against other established techniques. Classification testing is limited to subject and channel matching as those are the only labels. This shifts classification away from identifying specific waveforms toward a general similarity classification. Following from RA 1, each algorithm will undergo verification testing on all of the datasets. Then each datasets will be partitioned into fourths to allow for exhaustive LOOCV. The nature of I-Vectors requires a more nuanced approach to cross validation, shown in Table 3.1. The other algorithms will follow along by using only the testing and training data combinations while ignoring the enrollment data.

Results of all testing will produce EER portraying the sensitivity and specificity of each approach. Where applicable, the ordinal ranking of matches will also be compared to assess the clustering of each technique.

Table 3.1: The combinations of the split dataset (A, B, C, and D) to be used in the training and evaluaton of I-Vectors for the Principal Experiments. The example shows the number of variations for one distinct patchwork dataset in training, enrolling, and testing to produce a novel set of I-Vectors.

Iteration	Testing Data	Enrollment Data	Training Data
1	A	B	C,D
2	A	C	B,D
3	A	D	B,C
4	B	A	C,D
5	B	C	A,D
6	B	D	A,C
7	C	A	B,D
8	C	B	A,D
9	C	D	A,B
10	D	A	B,C
11	D	B	A,C
12	D	C	B,C

3.3.1 Principal Experiments: Algorithm Comparison

Comparing the performance of I-Vectors against other algorithms is necessary to set realistic goals for future development. These experiments will focus on the Partitioned Datasets which were thoroughly tested by the Core Experiments. However, this time data will be withheld from each algorithm for the purposes of creating training and testing sets. The alternative algorithms will be GMM-UBM, Gaussian Mixture Model based Hidden Markov Model (GMMHMM), SVMs, DP, and Siamese Neural Network (SNN) classifiers.

Each algorithm will report an EER related its classification ability. Where possible, likelihoods related to classification will be used to determine the confidence of classification and be used to generate clusters of similar matches. Performance will be evaluated on an experiment by experiment basis and as an averaged performance over the 24 experiments, for each training/testing and channel/subject combination. This results in 144 experiments which should take 10 weeks to complete.

3.3.2 Principal Experiments: Justification

Part of organizing LOOCV on the evaluation dataset is to enable other algorithms to classify the data given the unique sequence for producing I-Vectors. By sourcing results from multiple algorithms, edge cases of superior and inferior performance will provide a level of labeling otherwise unavailable. There are often instances where one algorithm will excel on a dataset all others struggle with, but this often indicates more about the data than the algorithms. Operating on unlabeled data increases the risk of encountering such a problem, but this is one way to at least identify problem datasets.

At the same time it provides a platform to benchmark performance of I-Vectors against approaches well known to those in the EEG community. Algorithms such as GMM-UBMs, GMMHMMs, and SVMs are commonly found in EEG literature. More novel techniques, like HDP and SNN, would also need to be vetted against these standardized tools. All of these algorithms would be evaluated by producing an EER and a likelihoods of classification.

3.4 RA 3: Driving Factors of I-Vector Performance

The results of RA 1 and RA 2 provide a framework for breaking down how I-Vectors classification operates on EEG based data. Drawing from the classification results of each algorithm, a *royalty* dataset consisting of the true negatives and positives and a *impostor* dataset consisting of the fasle negatives and positives.[2] It is likely that the royalty dataset will be large, so the most frequently occurring true negatives and positives will be used. Two *neighbor* datasets will be created based

[2]From the preliminary experiments I-Vector that exhibit royalty and impostor characteristics were discovered. Thus it is assumed they will exist in other datasets and can be leveraged to understand the characteristics of the data.

upon the closest sample, in terms of distance in the feature space of the common channel/subject, to act as a control for the royalty and impostor datasets.

The goal is find similarities and differences within and across each dataset that are responsible for the discrimination results of I-Vector and their counterparts. Each dataset will be evaluated for the distance between epochs and I-Vectors, distribution and influence of features, and likelihood of epoch occurrence based upon the determined distributions. Breaking down the statistical properties of the royalty and impostor epochs should give rise to relationships that explain the performance of the I-Vectors.

With such information it should be possible to tune a feature set that shifts, positively or negatively, the performance of I-Vector classification in a deterministic manner. If similar performance changes can be caused in the other algorithms this would suggest a fundamental understanding of the relationships present in the EEG datasets. Repeatedly performing the Principal Experiments is too time consuming so all tuned feature sets will operate on the combined royalty, impostor, and neighbor datasets called the *tracking* dataset. Thus each algorithm's performance can be tracked with respect to control and experimental data points, highlighting an understanding of the feature space. However, understanding the features alone is not sufficient to advocate for I-Vectors as an improved way to classify and cluster EEG datasets.

The final component of RA 3 focuses on why the dimensonality reduction and UBM to TVM transformation process best articulates these feature behaviors. Based upon the dataset distribution analysis, each individual dataset will be processed with ICA or PCA to identify the critical components of the features in each group (royalty, impostor, and neighbors) and as a whole. The tracking dataset will be evaluated following the protocol of the Core Experiments to provide baseline statistics for future comparisons.

Using the statistical patterns from the Core Experiments and the tracking dataset the performance of the tracking dataset will attempt to be controlled by inserting *linking* epochs. Given that royalty epochs exist, there should also be linking epochs that are able to extend and bridge the distributions inherent in the EEG data. While it is possible that some of the royalty epochs may be linking epochs, it is more likely that linking epochs will be equidistant to multiple classifications making them poor exemplars of classification. Their position as edge cases between feature distributions makes them integral to the development of I-Vectors as they create the decision surfaces driving compromise within the TVM.

3.4.1 Comparison Experiments: Epoch and Feature Impact

The Comparison Experiments are built from the aggregate results of the Core Experiments and Principal Experiments. For each algorithm a distribution the of true negatives/positives and false negatives/positives will be used to identify the severity of the incorrect classifications for each classifier and for the dataset at large. Each dataset (royalty, impostor, and neighbors) will be compared against various types of distributions (binomial, geometric, normal, etc.) using Kullback-Leibler Divergence (KLD) to characterize the behavior. Based upon the agreement to a normal distribution or multi-variate distribution PCA or ICA will then be used to isolate the dominant components of each dataset.

In conjunction with the properties of the datasets, their individual feature epochs, and paired I-Vectors, well be mapped with distance measurements to build a similarity mapping between samples. This is necessary to produce the neighbor datasets, but also for providing a reference point from which to compare experiments into features and dataset tuning. In the feature space distance between epochs will be found using ED and MD and CD for the I-Vectors. The relationship of how the UBM and TVM transform features into I-Vectors can then be tracked by these distance metrics. In

conjunction with the results of FA, it should then be possible to produce novel subsets of features to alter classification performance.

Together these experiments provide information to breakdown the impact of features on the development of I-Vectors. Testing the curated feature sets through the other algorithms should provide proof of understanding on the feature level. However, this assumes their performance can be controlled and aligns with the improvements seen in the I-Vectors. This would require re-running the Core Experiments with curated feature sets that caused statistically significant changes in the I-Vector classifications. It is hard to anticipate how diverse these feature sets will be, but 4 weeks should be ample time as only the datasets that produce the most and least false negatives and positives are used for each algorithm.

The final experiment involves constraining operations to the tracking dataset. By performing the Core Experiments on the tracking dataset and comparing the outcomes to the original results it should be possible to identify the different between the two datasets from a statistical perspective. It is assumed there will be a mismatch between the two, but that the addition of specific epochs, linking epochs, can align the feature space of the tracking dataset to that of the combine dataset. With the knowledge gained at this point, it should be possible to find linking epochs in the datasets or create them synthetically if they do not exist. The inclusion of specific epochs that improve classification and clustering of I-Vectors, but offer minimal improvement to other algorithms, would suggest the UBM and TVM transformation provides an intrinsic understanding of EEG data. These final experiments would likely take 6 weeks of work.

3.4.2 Comparative Experiments: Justification

At the heart of any ML technique are the features driving the discrimination and how those features impact results. As this work is using only cepstrum coefficients

the focus will be on which sets of these 26 features are driving the discrimination on the subject and subject-channel level. These specific features were shown to be effective when used by HMMs [153], but must now be understood for their impact on I-Vectors.

By sequentially removing features from the epoch data, UBM, and TVM it becomes possible to generate variations of I-Vectors from one training set. Starting with all 26 features and ending with a single feature will provide the influence each feature has on all the features remaining and all the features missing. Given enough resources it would be better to perform an exhaustive feature set search, but that is not practical. Despite this it may be necessary to perform something beyond sequentially removing features if there are indications that the features are not independent.

There is the possibility of producing synthetic features to prove the significance of linking epochs. This process was developed in the preliminary experiments for small feature sets, but lacked the necessary direction to be influential. The overlap of mixtures within the UBM is the ideal location for a linking epoch, however there is no assurance of a real epoch occupying this space. If such epochs could be readily found, it would most likely not be necessary to produce a TVM as the linking epochs could be converted into I-Vector providing classification thresholds.

3.5 Evaluation Metrics

The primary methods of evaluating epochs or their associated I-Vectors produce either a distance or a statistical likelihood. Depending on the classifier used there any likelihoods may be withheld leaving only a classification. This is part of the difficulty in evaluating ML algorithms for their ability to make decisions. However, I-Vectors do not have this problem as they provide a robust way to reduce the dimensionality of the data in a step-by-step process. The tools used to evaluate the behavior and discrimination ability of features and algorithms is presented here for completeness.

3.5.1 I-Vectors

The main evaluation metric for the I-Vector process are the distances between known and unknown data. When two I-Vectors represent the same segment of subject data, their CD is found to be 1. As the I-Vectors separate, being from the same subject, similar subjects, or unrelated subjects, the scores decrease eventually reaching -1. There are many factors that can impact the weights of the individual elements such as Within Class Covariance Normalization (WCCN) [91] and I-Vector length normalization [145]. Similar to how LDA is used to refine the dimensions of the I-Vectors from its TVM, these techniques attempt to reduce the dimesonality of the data.

Thus knowledge of the relationships between TVM elements is required to understand the resultant I-Vector. Otherwise, the CDs calculated are a weak measurement of the strength of the TVM's ability to translate epochs into I-Vectors. Therefore distance metric's based on I-Vectors are effective for subject verification and subject similarity, but fail to provide insight into the nature of the TVM. To fully track the development of the TVMs through these metrics it is necessary to tightly control the training data, which is part of the comparison experiments.

3.5.2 Epochs

Each epoch represents a point in feature space. As these should exist in the same space regardless of recording technique or subject, they can all be directly compared to each through distance calculations. These measures do not need to be perfect in terms of handling error or unequal feature weighting, but use of the MD should provide a balance to all of the EDs. What is most critical is that changes in the feature sets are propagate in a linear fashion through the distance calculations ensuring that at least relative change can be monitored. The intent is not to evaluate each feature as being a good or bad choice, but to understand how they influence each other and

the resultant I-Vectors. Thus any epoch evaluations should be universal in the even the EEG feature set changes.

The results of using epochs to train and evaluate each ML algorithm will be presented primarily as EER. Using the intersection of false negatives and false positives provides a conclusive way to compare the performance of each algorithm. This requires the data to be labeled in some manner so algorithms results may be compared to known results which is why the majority of work focuses on subject and subject-channel classification.

3.5.3 Distributions

A major component of this work relies on modeling the behavior of the datasets. Understanding the relationships between all of the datasets is necessary to isolate the influential distributions and features. For distributions the KLD, D_{KL}, can be found as an indication of similarity between the discrete datasets.

$$D_{KL}(P||Q) = -\sum_{i} P(i) \log \frac{Q(i)}{P(i)} = \sum_{i} P(i) \log \frac{P(i)}{Q(i)} \tag{3.5-1}$$

The discrete version would operate directly on the samples in the datasets, carried out during the Comparison Experiments. The continuous version

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \tag{3.5-2}$$

is setup to address divergence between the mixtures of each UBM given the presence of means and variances. Together they provide a direct metric for quantifying how well subsets of the datasets and mixtures represent their parent distributions.

This becomes increasingly important for the royalty and impostor datasets in determining their likelihoods. The likelihood of these epochs occurring based upon the known dataset distributions and UBMs provides insight into their nature. If epochs in

the impostor dataset are determined to be extremely rare, their classification may not matter because they could artifacts. Likewise, if the royalty epochs show equivalent likelihood from all datasets they may represent background. Therefor it is necessary to know the underlying distributions to characterize the individual features and the epochs themselves.

CHAPTER 4

Preliminary Reseaerch

I-Vectors were developed for use by the speech recognition community. The seminal papers concerning their methodology and use focus on speech data which is reasonably similar to EEG data in terms of being band limited non-stationary signals. However, EEGs are produced by multiple independent sources and range their waveforms are not as well defined as speech phonemes. These differences are not insurmountable, as evidenced by the number of ML algorithms shared between the two fields. This simply requires preliminary experiments to understand how to tune the system for a different type of data.

There are few parameters to control when producing I-Vectors, but each impacts a distinct part of the process. When producing the UBM a mixture size must be specified. The UBM is then used to develop the TVM through a given number of EM iterations. Finally, a length for the I-Vectors must be chosen, although this has an upper bound set by the data. The most important parameter is how the training, enrollment, and testing datasets are constructed and partitioned.

Each parameter must be tested individually to isolate changes in performance. The preliminary experiments address performance on different discrimination levels (subjects, trial, and channel) as a function of UBM mixture size. Using a synthetic dataset modeled on the TUH Corpus and the diverse PhysioNet datasets provides variation in data itself related to number of subjects, duration of recordings, and recording content. The experiments themselves are based on subject verification where the training, enrollment, and testing data are identical.

These loosely defined experiments provided a platform to explore the ability of I-Vectors to discriminate EEGs. The results were critical for exploring how I-Vectors work in a general sense, but also specifically for the new medium of EEGs. These preliminary results provided insight in how to propose experiments to address the research aims. A discussion of these factors is provide to show how these insights will be applied moving forward.

4.1 Preliminary Research

By performing experiments with identical training, enrollment, and testing data strong *primary classification* performance is assured. Primary classification is the ability of I-Vectors to match on targeted level of data hierarchy. When comparing channel I-Vectors, this would be seen by each channel I-Vector matching best with itself. *Secondary classification* would be how well each subject's channel I-Vectors match against the other channel I-Vectors from the same subject.

I-Vectors are primarily evaluated by comparing the CD between them. This allows for effective verification testing, but also provides a strong clustering mechanic enabling secondary classification. The performance of primary classification is hypothesized to be near perfect given the nature of the datasets. While secondary classification is assumed to perform poorly given the complexity of the data and that this is the first time such work has been attempted.

To help with the transition to EEG data, idealized datasets are used to control for problems directly related to data quantity and quality. A simplified version of a subset of the TUH Corpus is synthesized as a proof of concept test-bed. This *Synthetic Dataset* allows the number of underlying mixtures to be controlled and is described in the ensuing section. Controlling the composition of the data helps test the impact of varying the UBM mixture size on classification.

In addition, EEG recordings typically contain an abundance of *background* epochs. These are portions of data with waveforms deemed insignificant to the study. This contrasts with speech data which is typically comprised only of waveforms of interest. The PhysioNet dataset contains a balance of background and subject activity given the directed experiment protocol. This makes it similar to the structure of speech data than a typically 20+ minute EEG recording full of benign waveforms.

4.1.1 Synthetic Dataset

The synthetic datasets were created to isolate the impact of (1) a common feature across subjects, and (2) a unique feature for each subject. The three synthetic datasets were based on raw data from the TUH Corpus, and are referred to as *simulated*, *static* (simulated with an additional common feature across subjects) , and *unique* (simulated with a unique feature for each subject). Figure 4.1 provides a representation of the data generation process.

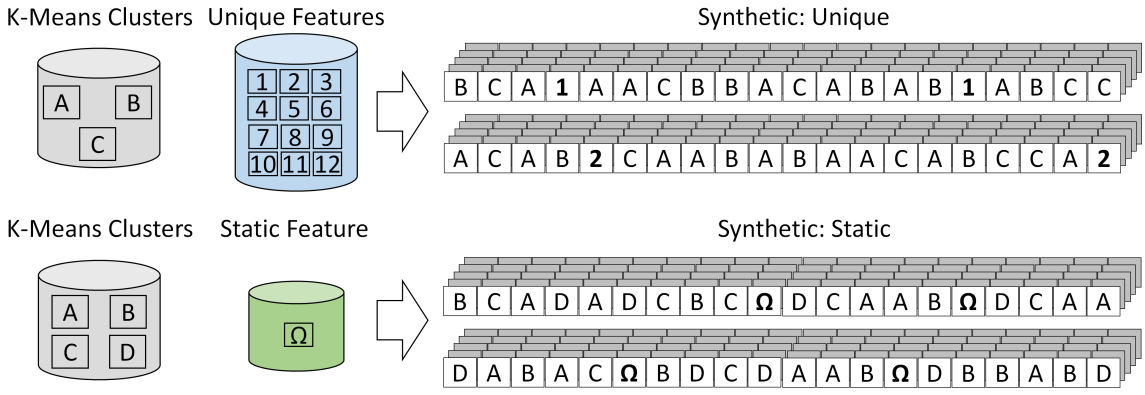

Figure 4.1: Generation of synthetic data from the TUH Corpus. The clustered K-Means data (gray) and the unique (blue) or static (green) features enable the creation of unique and static synthetic datasets. Only 10% of the subject's K-Means generated data is replaced by the external feature. The models produce all 22 channels at once for each of the 12 simulated TUH Corpus subjects.

The simulated data used K-Means on 12 subjects to produce Gaussian models with K of size 3 and 4. By grouping all 22 channels for each epoch, the 27 cepstrum

features per channel were turned into a cluster of 594 linked means and variances. The likelihood of each cluster was used to the drive the continuous HMM that produced the resultant models. This process created Simulated 3GMM and 4GMM datasets containing 12 subjects each with 22 channels with a duration of 10 minutes.

The static and unique datasets start as simulated datasets which then gain an additional feature. In the case of the static datasets, the additional feature was a randomly chosen Gaussian mixture from a 16 mixture UBM generated from a random subject within the PhysioNet dataset. The unique datasets drew from the same UBM, but each was randomly assigned one of the top 10 weighted Gaussian distributions to draw from. In both cases a random 10% of the base simulated dataset was replaced with the new feature.

This produces six unique synthetic datasets: Sim3, Sim4, Sta3, Sta4, Uni3, Uni4, outlined in Figure 4.1. As the static and unique datasets contain an additional feature the data sets model mixtures of size 3, 4, and 5. As much authenticity of the raw data is preserved in the synthetic data, highlighted in Table 4.1. The feature sampling rate remains 10Hz. Only the duration of the recordings changes, held constant at 10 minutes. This is longer than the 2 minute recordings of PhysioNet and shorter than the 20+ minute recordings of TUH Corpus.

Table 4.1: Composition of Synthetic TUH Corpus Datasets

Name	Type	Features	Channels	Duration
Sim3	Simulated	3	22	600s
Sta3	Static	4	22	600s
Uni3	Unique	4	22	600s
Sim4	Simulated	4	22	600s
Sta4	Static	5	22	600s
Uni4	Unique	5	22	600s

4.1.2 Experiments

The experiments consist of 'target' verification, with the targets being subject, trial, and channel. Subject verification of the synthetic dataset requires discriminating 12 unique subjects. The PhysioNet dataset allows for verification of the 109 subjects, 1,526 trials, and 33,572 channels. These experiments represent the four primary classifications of kind-to-kind matching, how well the system matches the enrollment I-Vector to the testing I-Vector. Aside from the channel I-Vector performance, these results are presented as functions of the UBM mixture size.

The secondary classifications evaluate the trials and channels of the PhysioNet dataset their associated cohorts. For trial I-Vectors the cohorts are built from the trials of the given subject and trials of the same experimental protocol [1]. The channel I-Vector cohorts are built in the same manner, around the subject and around their common trials. These secondary classifications do not use a range of UBM mixture sizes, but instead of single mixture size shared among the experiments.

Performance for these experiments is measured in terms of sensitivity and specificity. For the primary classifications this is displayed as an EER. When working with the secondary classifications, performance is measured in terms of being within the cohort set. When comparing channels to a given subject or trial the within set consists of the top 22 matching channels. Any matches that are of type-for-type (a channel matching to a native subject or trial) are labeled as verification matches, kind-for-kind matches (channel matches into a different subject-trial) are primary, and matches across non-subject-trials are secondary matches. Classification via GMM-UBMs is also performed fro the primary classification group with results presented as EERs.

[1] Recall that each subject performed 14 trials of which 12 were a repeated set of 4 common sequences of activity. Thus each subject has 4 sets of repeated trials and these trial protocols were common for all subjects.

Synthetic TUH Corpus Testing Results

Given the controlled nature of the synthetic data and the verification testing protocol, the major factor controlling performance is the UBM mixture size. Classification performance of UBMs and I-Vectors as a function of UBM mixture size is presented in Figures 4.2 and 4.3. The original and synthetic datasets classifications using GMM-UBM exhibit improved performance as the mixture size increases. The rate of EER reduction in the two unique datasets (Unique 3GMM and Unique 4GMM) is the strongest and most responsive reaching zero with 32 mixtures. The remaining datasets reduce their EER by roughly 10% over the 10 mixture sizes.

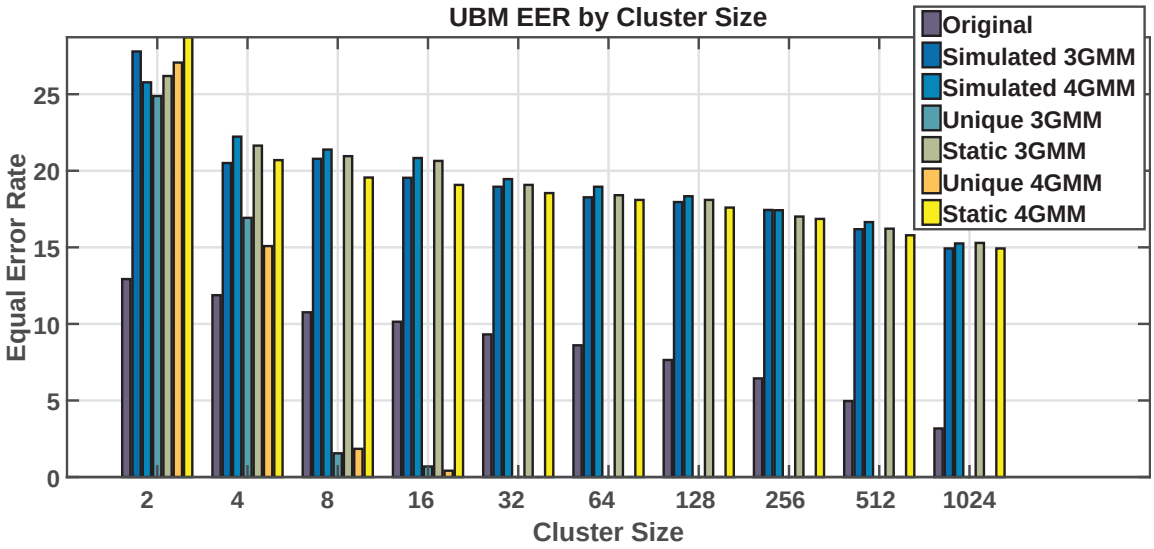

Figure 4.2: EER of UBMs on the seven data sets (l to r) Original, Simulated 3GMM, Simulated 4GMM, Unique 3GMM, Static 3GMM, Unique 4GMM, Static 4GMM. The EER for two unique data sets reaches 0% when the models exceed 16 clusters.

Evaluating the performance of I-Vectors as a discriminators, figure 4.3, shows that nearly all datsets achieve a near zero EER with 4 UBM mixtures. The two exceptions, Static 3GMM and Static 4GMM, require a mixture size of 8 to reach a near zero EER. Unlike their UBM based counterpart, none of the datasets ever settle to an EER of zero.

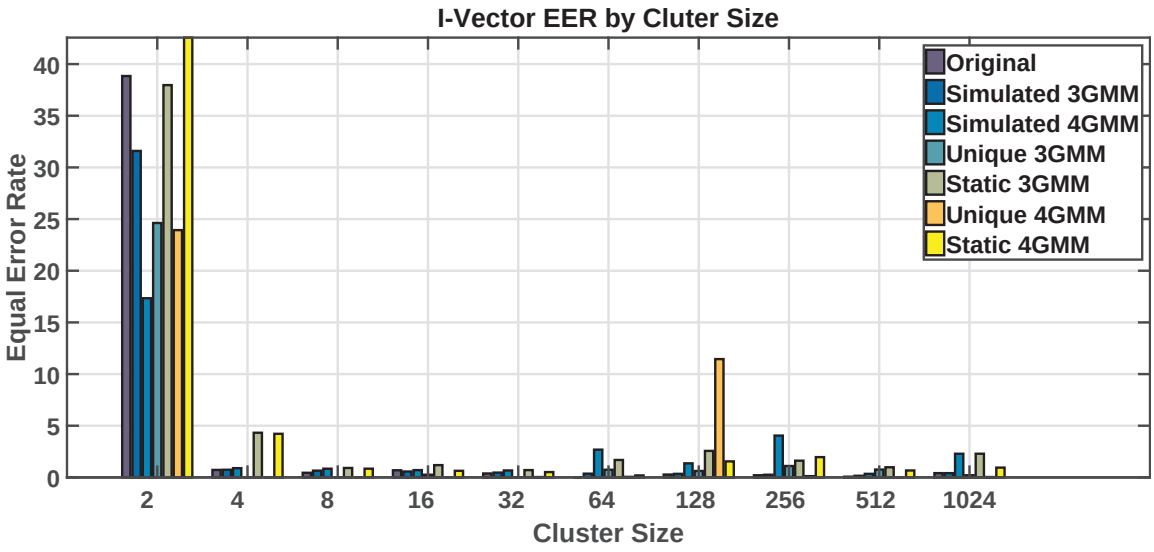

Figure 4.3: EER of I-Vectors on the seven datasets (l to r) Original, Simulated 3GMM, Simulated 4GMM, Unique 3GMM, Static 3GMM, Unique 4GMM, Static 4GMM. A strong reduction in EER is seen when transitioning from 2 to 4 clusters for modeling. Beyond this transition changes to the EERs at higher cluster sizes are minimal.

PhysioNet Testing Results

Verification testing on the subject and trial I-Vectors produced the EEG seen in Figure 4.4. At lower mixture sizes, the I-Vector based classification methods outperform their GMM-UBM counterparts. This trend does not continue with the trial I-Vector classification which performs worse than the GMM-UBM classifier at larger mixture sizes. Subject GMM-UBM performance appears to plateau at larger mixture sizes, similar to subject I-Vector performance. The trend of increasing mixture size improving performance, noted in Figure 4.3, is seen in each classification aside from the trial I-Vectors.

The variance in primary classification using the trial I-Vectors of a given subject, separated between Full Trials and Motion Trials, is seen in Figure 4.5. Matching the 14 Full Trials or 12 Motion Trials I-Vectors within a subject correctly plateaus with at a UBM with 8 mixtures, while equivalent performance of GMM-UBMs requires 256 mixtures in the UBM. However, beyond 256 mixtures the GMM-UBM classification

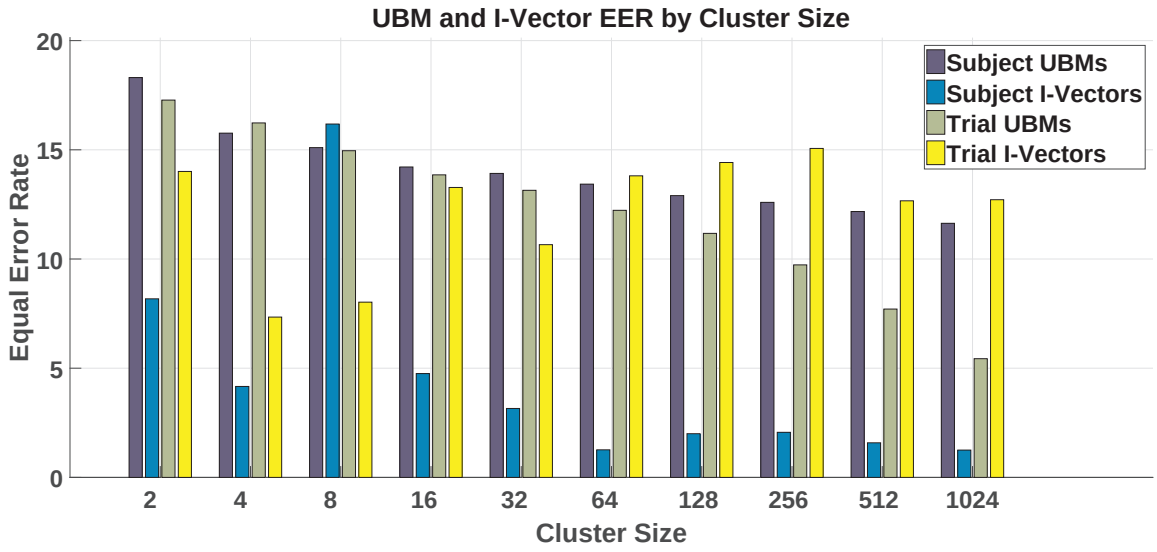

Figure 4.4: PhysioNet UBM and I-Vector verification test results as a function of cluster size. The leftmost bars represent Subject UBMs and I-Vectors and the rightmost bars represent Trial UBMs and I-Vectors.

produces a lower EER and a smaller variance than the I-Vectors. Classification of Full Trials versus Motion Trials for the I-Vectors produces equivalent performance, but for GMM-UBMs Full Trials' EER the mean is always one standard of deviation stronger than the Motion Trials' EERs.

Within each subject are 14 trials, 12 are motion based trials and 2 are resting calibration. These can be broken into their respective groups, Figure 2.3, and classified based upon how well they match into various trial sets, Figure 4.3. A breakdown of the groupings is given in Table 4.2. Evaluating the secondary clustering classifications against their expected likelihood distribution, Table 4.4, shows performance of the secondary classification.

In all instances the trial I-Vectors the expected likelihoods for producing complete sets of matches is statistically significant. This is most noticeable in the reduction of individual matches and the increase in full sets, 3 of 3 and 5 of 5. Allowing for the resting states has a strong impact on increasing the number of matches, given the reduction in single matches from 65.45% for Motion 3 of 3 to 17.62% for Full 5 of 5.

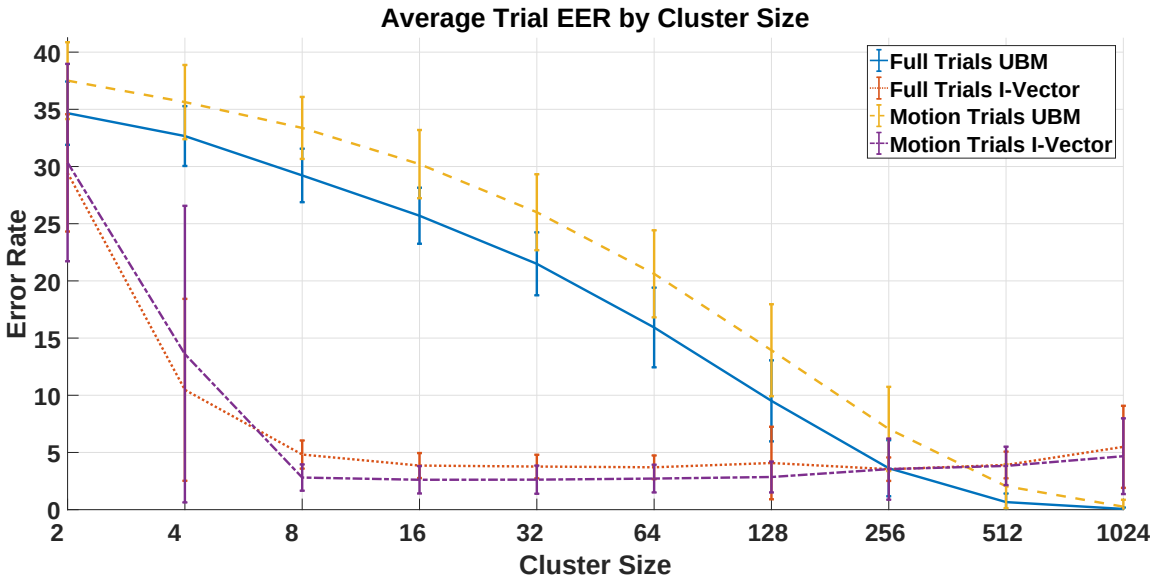

Figure 4.5: The averaged EERs of each PhysioNet subjects' trials as a function of cluster size. Error bars represent +/- one standard deviation across the entire subject set. The I-Vectors' EER plateaus after 8 clusters, while the UBMs' EER decreases as the cluster size grows.

Table 4.2: The Data Group specifies the trials given for the search space. The Cohort Groups show which trials are considered a distinct group. The Search Interval defines the acceptable positions [a-b] out of the available trials presented in the Data Group.

Label	Data Group	Cohort Groups	Search Interval
Motion 3/3	G1, G2, G3, G4	{G1}{G2} {G3}{G4}	[1-3] of 12
Full 3/3	G0, G1, G2, G3, G4	{G1}{G2} {G3}{G4}	[1-3] of 14
Full 3/5	G0, G1 G2, G3, G4	{G1}{G2} {G3}{G4}	[1-5] of 14
Full 5/5	G0, G1 G2, G3, G4	{G0 G1}{G0 G2} {G0 G3}{G0 G4}	[1-5] of 14

The doubling of Full 5 of 5 from an expected 2 matches to 4 matches out of 1308 trial sets.

The final secondary classification focused on channel I-Vectors fitting into their subject and trial cohorts. In this context there are three potential cohorts: a *verification* cohort - matching into the correct subject or subject-trial, a *primary* cohort -

Table 4.3: Experimental Likelihoods of Cohorts Using Four Clusters. Values indexed with * are significant at p<0.005 and with ˆ are significant at p<0.001

	1	2	3	4	5
Motion 3 of 3	55.66ˆ	41.13ˆ	3.21*	-	-
Full 3 of 3	66.44ˆ	33.72ˆ	2.14*	-	-
Full 3 of 5	38.00ˆ	51.07ˆ	10.93ˆ	-	-
Full 5 of 5	15.29ˆ	43.88ˆ	33.64ˆ	6.88ˆ	0.31*

Table 4.4: The expected likelihoods for cohorts of PhysioNet trials. The probabilities are generated from *p choose n* using the parameters set forth in Table 4.3

	1	2	3	4	5
Motion 3 of 3	65.45	32.73	1.82	-	-
Full 3 of 3	70.51	28.21	1.28	-	-
Full 3 of 5	49.45	24.73	2.75	-	-
Full 5 of 5	17.62	47.00	30.18	5.04	0.14

channel-trial matches occupying the top 14 matches of a trial from the native subject, and a *secondary* cohort - channel-trial matches occupying the top 14 matches of a trial from non-native subjects. Their aggregation is represented by the *total* cohort. These match percentages are presented in Table 4.5.

Table 4.5: Subject and Trial Cohort Matching Using Eight Clusters. The Channel I-Vector's distances are averaged and ordered to produce trial and subject matches. Verification shows the likelihood the top match pairs with the native trial or subject. Primary shows the likelihood of matching to any of the 22 channels trials within the trial group in the trial or subject. Secondary shows the likelihood of matching to trials within the trial group among all other subjects. These results are statistically significant.

Data	Verification	Primary	Secondary	Total
Sub to Chan	76.15	-	-	-
Chan to Sub	82.57	-	-	-
Trial to Chan	81.06	52.43	9.48	61.91
Chan to Trial	96.66	63.23	7.61	70.84

In this parent (subject/trial) to child (channel) relationship, the verification cohort child-to-parent relationship appears stronger (82.57% and 96.66%) than the parent-to-child relationship (76.15% and 81.06%). The primary cohort results indicate over 50% of the matches originate from the native subject-trial. Included the secondary cohort, over 60% of trial-to-channel and channel-to-trial matches are relevant to the subject or trial under test.

4.2 Preliminary Results Discussion

These experiments served to prove the feasibility of I-Vectors on a new medium and achieved that in the discrimination of subjects, trials, and channels. While the initial experiments are straightforward subject verification testing, the performance of the secondary classifications suggests the I-Vectors are capturing nuance in the data beyond the primary classification. The experiments are a success because of the strength of the primary and secondary classifications. Most interesting is the statistical significance the trial groupings seen in Table 4.3.

4.2.1 Primary Classifications

The results of the synthetic classification, Figure 4.3, illustrate the strength of I-Vectors when the system utilizes a UBM with mixtures equal to or exceeding those found in the data. This is expected as it is the foundation of the technique, however the non-zero EERs at larger mixture sizes is not expected. As the dataset is tightly controlled and only the number of mixtures is changing with each evaluation there are two possible sources causing this behavior. The first would be that the UBMs are failing to accurately represent the dataset as they grow in size and the other is that the process of generating the TVM fails to produce a matrix of sufficient quality.

A failure of the UBM to represent the dataset should be seen in EER of GMM-UBM and I-Vector. Reviewing Figure 4.2, the GMM-UBM EERs shows that, while not as

strong as the I-Vectors, their performance does not worsen as the UBM mixture increases. This suggests the reason for the variation in performance is due to the development of the TVM. Prior testing showed that given a TVM the creation of I-Vectors is a deterministic process, but it was not considered necessary to analyze the production of the TVM.

This conclusion is troubling because it casts doubt on the other preliminary results. A poorly formed TVM will negatively impact the ability of the resultant I-Vectors to discriminate. In fact, this could be the reason for poor performance of the trial I-Vectors seen in Figure 4.4. Classification using GMM-UBMs was only included to show the performance of the base classification offered by the UBM so any I-Vector performance worse than it would suggest something is wrong with the I-Vectors. This could also explain why, despite the GMM-UBMs being able to achieve near zero EER for trial classification in Figure 4.5, the I-Vectors could not reach equivalent performance.

As the TVM is core mechanic through which the I-Vectors operate, it is necessary to understand why the training process may be insufficient for EEG based data. From the preliminary experiments it is unclear if the issue is related to the organizational structure of training data (subjects, trials, and channels) or perhaps the variability of the waveforms in the data. The I-Vector variances in Figure 4.5 suggests the occurrence of insufficient TVMs may be rare or is not as pronounced with small datasets.

These concerns should be addressed by articulating specific datasets, where the variability of subjects, trials, and channels can be controlled with respect to variance and quantity. In doing so the UBM can manipulated with its properties being tracked through the EM training process of the TVM. If near-identical UBMs can be produced from different datasets or subsets of a larger dataset, akin to a DP, it may be possible to produce an archetypal UBM that covers the full dataset.

4.2.2 Secondary Classifications

The secondary classifications presented in Tables 4.3 and 4.5 suggest I-Vectors can identify facets of multiple discrimination levels (subject, trial, and channel) in EEG data. This is promising, but expected given the success of the technique within the speech recognition community. The difficulty in leveraging it with EEG data is that a pathway to link the resultant I-Vectors to the underlying EEG characteristics is missing. Speech produces I-Vectors on much smaller time scales than those used in the preliminary experiments which may not be feasible for EEG data.

Developing synthetic datasets with purposefully mismatched channels or predetermined phenomena spliced across subjects and channels should allow for detailed analysis of what the I-Vector process deems significant for discrimination. Such a process would not be reliant on annotated data so long as subject differences are strong enough to be discernible, which is strongly suggested by the secondary classification results. In this manner, subject data could be stitched together in various segment lengths. The purpose of this would be to (a) test how a variable duration 'trials' impact I-Vector generation and (b) allow for better control when performing similar secondary classification experiments with unlabeled data by using the subjects as the labels.

APPENDIX A

Equations

A.1 Features

A.2 Algorithms

A.3 Evaluation Metrics

APPENDIX B

Graphics

B.1 Test Configurations

B.2 Ideal Results

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] 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.

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

[4] B. J. Lance *et al.*, "Brain-computer interface technologies in the coming decades," *Proc. IEEE*, vol. 100, no. SPL CONTENT, pp. 1585–1599, 2012.

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

[6] 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.

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

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

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] B. J. Lance *et al.*, "Brain Computer Interface Technologies in the Coming Decades," 2012.

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

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

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

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

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

[71] 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.

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

[73] 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.

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

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

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

[77] E. Baar and B. Güntekin, "A review of brain oscillations in cognitive disorders and the role of neurotransmitters," *Brain Res.*, vol. 1235, pp. 172–193, 2008.

[78] 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.

[79] 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.

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

[81] 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.

[82] 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.

[83] 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.

[84] S. Fazli, M. Danóczy, J. Schelldorfer, and K.-R. Müller, "1-penalized linear mixed-effects models for high dimensional data with application to BCI," *Neuroimage*, vol. 56, no. 4, pp. 2100–2108, 2011.

[85] 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.

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

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2529310," *Society*, vol. 33, no. 1, pp. 159–174, 2008.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

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

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

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

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

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

[104] 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.

[105] 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.

[106] 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.

[107] 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.

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

[109] 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.

[110] 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, 2012.

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

[112] 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.

[113] 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.

[114] 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.

[115] J. C. Sackellares *et al.*, "Quantitative EEG analysis for automated detection of nonconvulsive seizures in intensive care units." *Epilepsy Behav.*, vol. 22 Suppl 1, no. 0 1, pp. S69–73, 2011.

[116] 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.

[117] 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.

[118] 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.

[119] 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.

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

[121] 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.

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

[123] 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.

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

[125] S. Marcel and J. D. R. Millán, "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.

[126] 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.

[127] 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.

[128] 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.

[129] 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.

[130] 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.

[131] 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.

[132] 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.

[133] 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.

[134] 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.

[135] 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.

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

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

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

[139] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, may 2007.

[140] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice Modeling With Sparse Training Data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, may 2005.

[141] 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.

[142] 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.

[143] 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.

[144] 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.

[145] 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.

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

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

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

[149] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. January, p. Appendix 3A, 1986.

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

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

[152] 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.

[153] 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.