Automated detection of interictal epileptiform discharges in scalp electroencephalogram of patients with epilepsy

Bagheri, Elham

2019

Bagheri, E. (2019). Automated detection of interictal epileptiform discharges in scalp electroencephalogram of patients with epilepsy. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/89750>

<https://doi.org/10.32657/10220/48049>

Automated Detection of Interictal Epileptiform Discharges in Scalp Electroencephalogram of Patients with Epilepsy

ELHAM BAGHERI

SCHOOL OF ELECTRICAL & ELECTRONIC ENGINEERING

2019

**AUTOMATED DETECTION OF INTERICTAL
EPILEPTIFORM DISCHARGES IN SCALP
ELECTROENCEPHALOGRAPH OF PATIENTS
WITH EPILEPSY**

ELHAM BAGHERI

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

...09-April-2019...

Date

ELHAM BAGHERI

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

...09-April-2019...

Date

.....

Assoc. Prof. JUSTIN DAUWELS

Authorship Attribution Statement

This thesis contains material from three papers published in the following peer-reviewed journals and papers accepted at conferences in which I am listed as an author.

Preliminary results of Chapter 3 are published as "**E. Bagheri**, J. Jin, J. Dauwels, S. Cash, and M. B. Westover, "Fast and efficient rejection of background waveforms in interictal EEG," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 744-748." We are preparing a journal paper publication from the final methods and results as well.

The contributions of the authors are as follows:

- Prof. Justin Dauwels co-designed the study, provided the initial project direction, guided throughout the project, and helped in revising the manuscript drafts.
- I co-designed the study, preprocessed the dataset, developed and evaluated the algorithms, and prepared the manuscript. I performed all the data analysis, coding and algorithm design at the School of Electrical & Electronic Engineering.
- Dr. Jing Jin helped in collecting the dataset, provided feedback on the project, and helped in revising the manuscript.
- Dr. M Brandon Westover and Dr. Sydney S. Cash provided and annotated the dataset, provided feedback on the study and clinical matters, and helped

in revising the manuscript.

Chapter 4 is published as **E. Bagheri**, J. Jin, J. Dauwels, S. Cash, and M. B. Westover, "Classifier cascade to aid in detection of epileptiform transients in interictal EEG," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 970-974.

The contributions of the authors are as follows:

- Prof. Justin Dauwels co-designed the study, provided the initial project direction, provided guidance throughout the project, and helped in revising the manuscript drafts.
- I co-designed the study, preprocessed the dataset, developed and evaluated the algorithms, and prepared the manuscript. I performed all the data analysis, coding and algorithm design at the School of Electrical & Electronic Engineering.
- Dr. Jing Jin helped in collecting the dataset and provided feedback on the project.
- Dr. M Brandon Westover and Dr. Sydney S. Cash provided and annotated the dataset, provided feedback on the study and clinical matters, and helped in revising the manuscript.

Chapter 5 is published as **E. Bagheri**, J. Dauwels, B. C. Dean, C. G. Waters, M. B. Westover, and J. J. Halford, "Interictal epileptiform discharge characteristics underlying expert interrater agreement," *Clinical Neurophysiology*, vol. 128, no.

10, pp. 1994-2005, 2017.

The contributions of the authors are as follows:

- Prof. Justin Dauwels co-designed the study, provided the initial project direction, guided throughout the project, and helped in revising the manuscript.
- I co-designed the study, preprocessed the dataset, developed and evaluated the algorithms, and prepared the manuscript. I performed all the data analysis, coding and algorithm design at the School of Electrical & Electronic Engineering.
- Dr. Brian C. Dean and Dr. Chad G. Waters helped in preparing the dataset.
- Dr. M Brandon Westover provided feedback and guidance on the study and clinical matters, and helped in revising the manuscript.
- Dr. Jonathan J. Halford provided the dataset, co-designed the study, provided guidance throughout the project and clinical matters, and helped in revising the manuscript.

09-April-2019

.....

Date

ELHAM BAGHERI

Acknowledgements

I would firstly like to express my sincere gratitude to my supervisor Prof. Justin Dauwels for supporting my PhD study and research; and also for his patience, great guidance and motivation, in addition to his emotional support throughout my PhD study and preparing this thesis. His immense knowledge, insightful and constructive comments help me gain knowledge in my research area and motivate me to explore new ideas. In addition, our regular meetings encouraged me to learn more and gain constant feedback about my research.

My thesis work was in collaboration with Medical University of South Carolina (MUSC), and Massachusetts General Hospital (MGH) and Harvard Medical School.

I would like to thank Dr. Brandon Westover, and Dr. Sydney Cash from MGH, who provided the annotated clinical EEG dataset required for my research study. Moreover, our in-person and online meetings and discussions provide constructive feedback regarding my research, and help me gain deeper knowledge and widen my research from various aspects.

My sincere thanks also goes to Dr. Jonathan J. Halford from MUSC. I wish to thank him for providing a dataset of clinical EEG, which is essential for my research. He gave me many ideas and motivation during our online discussions and meetings, and always provided me with instant and comprehensive feedback and suggestions; I thank him for all his support.

Besides, I would like to thank my friend and colleague, Dr. Jing Jin. She supported me from very beginning of my research studies. She explained

concepts patiently to me, and helped in dealing with difficulties that I faced during initial stages of my research. I also thank all my friends and colleagues for their technical and emotional support and fruitful discussions.

My thanks also go to Prof. Lin Zhiping from school of Electrical and Electronic Engineering, and Prof. Suresh Sundaram from school of Computer Engineering, who are the members of my thesis advisory committee. I wish to thank both of them for their insightful comments, encouragement, and providing research materials and ideas.

Finally, I wish to thank my father, Arman Khaghani and my mother, Elah Ebrahimi for devotedly supporting me through all aspects of my life and studies. Most importantly, I am grateful to God Almighty for all blessings.

Table of Contents

Table of Contents	
Summary	i
List of Figures	iii
List of Tables	ix
List of Acronyms	xi
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	5
1.3 Major Contributions of the Thesis	8
1.4 Organization of the Thesis	10
2 Literature Survey	11
2.1 Definition of Terms	11
2.1.1 Epilepsy	11
2.1.2 Electroencephalography	13
2.1.3 Interictal epileptiform discharges (Spikes)	14
2.2 IED Detection Methods	15
2.2.1 Template matching	16
2.2.2 Parametric or autoregressive models	17
2.2.3 Mimetic analysis	19

2.2.4	Power spectral analysis	22
2.2.5	Wavelet analysis	24
2.2.6	Artificial neural networks	26
2.2.7	Deep learning	31
2.2.8	Artifact rejection	35
2.3	Comparison of IED detection methods	41
2.4	Inter-rater agreement analysis methods	49
3	A Fast Automated Method to Facilitate the Detection of Interictal Epileptiform Discharges in the Scalp Electroencephalogram	51
3.1	Introduction	51
3.2	Materials and Methods	52
3.2.1	Scalp EEG data	52
3.2.2	Data preparation and feature extraction	53
3.2.3	Implementing the background rejection algorithm	59
3.3	Results	64
3.4	Discussion	73
3.5	Conclusions	75
4	Classifier Cascade to Aid in Detection of Epileptiform Transients in Interictal EEG	77
4.1	Introduction	77
4.2	Materials and Methods	78
4.2.1	EEG dataset	78
4.2.2	Classifier cascade	79
4.2.3	Combining the cascade with an IED detection method . .	81
4.3	Results and Discussion	82
4.4	Conclusions	87

5	Interictal Epileptiform Discharge Characteristics Underlying Expert Interrater Agreement	89
5.1	Introduction	89
5.2	Materials and Methods	90
5.2.1	EEG scoring and EEGnet software	90
5.2.2	EEG dataset description	90
5.2.3	Scoring the EEG dataset by experts	91
5.2.4	Inter-rater agreement analysis	92
5.2.5	Signal analysis methods	95
5.3	Results and Discussion	106
5.3.1	Expert annotation results	107
5.3.2	EEG waveforms	109
5.3.3	Visualization by t-SNE	111
5.3.4	Expert scoring results	113
5.3.5	Features versus expert scoring	116
5.3.6	Electrode position on scalp	118
5.3.7	Feature selection	119
5.3.8	Multivariate analysis of expert agreement	124
5.4	Conclusions	126
6	Computers vs Human Electroencephalographers for Recognizing Epileptiform Discharges	132
6.1	Introduction	132
6.1.1	Persyst IED detection software	134
6.2	Materials and Methods	136
6.2.1	Dataset of EEG recordings and expert opinion	136
6.2.2	IRA using Gwet's AC2	138
6.2.3	Receiver operating characteristics (ROC) based on expert consensus	138
6.2.4	Mean pairwise ROC analysis	139

6.2.5	Bootstrap statistical analysis of mean pairwise ROC performance	139
6.3	Results and Discussion	142
6.3.1	Inter-rater agreement analysis	142
6.3.2	ROC based on experts' consensus	143
6.3.3	Mean pairwise ROC analysis	144
6.3.4	Bootstrap statistical analysis of mean pairwise ROC . . .	144
6.4	Conclusions	152
7	Conclusion and Recommendations	155
7.1	Background Rejection Algorithm	155
7.2	Classifier Cascade	157
7.3	Inter-rater Agreement Analysis in IED Detection	158
7.4	Computers versus Humans	160
	Author's Publications	162
	Bibliography	163

Summary

Finding interictal epileptiform discharges (IED) or spikes in the electroencephalogram (EEG) is a part of diagnosing epilepsy. Automated annotation of interictal EEG is important for epilepsy diagnosis and management. Manual annotation of IEDs is tedious and there is substantial disagreement among experts. In the present thesis, various methods contributing to automated IED detection are proposed and investigated.

We develop a feature-based method, in the form of a cascade of simple fast classifiers to reject most background waveforms and to speed up IED detection. We can eliminate 78.41% of background waveforms while retaining 96.93% of IEDs on our cross-validated dataset comprising 156 subjects. Rejecting background by our proposed method, speeds up the classification by a factor ranging from 4.69 to 1.76 for the considered classifiers.

Next, we propose an ensemble algorithm comprising a cascade of SVM classifiers to enhance the precision of IED detection. At a fixed sensitivity, we are able to improve precision by 6.78%; and at a fixed false positive rate, the sensitivity improves by 2.83%.

Additionally, we utilize a dataset annotated by several neurologists to analyze inter-rater agreement and signal characteristics. We determine features correlating with expert scoring and contributing to agreement or disagreement among experts, and fit models to predict expert scores. The features that best predict inter-rater agreement among experts regarding the presence of IED are certain wavelet features, using locally standardized EEG. Our models for pre-

dicting expert opinion based on EEGer's scores perform best with a large group of EEGers (more than 10).

Finally, we develop and implement several statistical measures for evaluating the performance of a IED detector and comparing it with human experts. These measures include the mean pairwise sensitivity and false positive rate performance of the P13 algorithm versus 19 academic EEGers, and a statistical Turing test using accelerated bootstrap analysis.

List of Figures

1.1	Epileptiform spike (IED) in one EEG recording from MUSC dataset.	2
1.2	Examples of different varieties of IED morphology from MGH dataset.	3
2.1	Linear prediction model used by Sankar and Natour [41].	20
2.2	Parameters extracted from mimetic analysis [53].	21
2.3	Feature points and reference diagram for feature extraction [61].	23
2.4	The two-dimensional array of neurons in the self-organizing fea- ture maps (SOFMs).[53].	28
2.5	Architecture of the radial basis function network (RBFN) used by Acir <i>et al.</i> [62].	30
2.6	The feed forward artificial neural network (ANN) using error back propagation, by Carey <i>et al.</i> [85].	31
2.7	Structure of the convolutional neural network [91].	32
2.8	A network with eight input and four output units, and two mem- ory cell blocks of size 2 [87].	32
2.9	A typical DBN which contains 2 Belief Nets and 2 RBMs [99].	34
2.10	Example of the 3 independent components extracted in the train- ing phase after reducing dimension to 3 using PCA [117]. (TF = time frames)	38
2.11	Clustering results of the dipole position and orientation vectors representing focal brain activity [118].	40

3.1	Subband decomposition for implementation of discrete wavelet transform.	57
3.2	Wavelet decomposition of a sample IED.	57
3.3	Flowcharts showing (a) training and (b) testing processes.	59
3.4	The features selection process applied in each step by computing the empirical cumulative density function.	61
3.5	Two procedures for IED detection (a) using a classifier and (b) applying the background rejection algorithm followed by a classifier.	63
3.6	The feature $DWT-rbio3.3-\max(a_2)$, computed for a sample EEG recording. The horizontal line shows the threshold to be applied on it.	65
3.7	Overall percentage of rejected background after every step, for training and testing sets.	68
3.8	Sensitivity and specificity versus the number of IEDs in each EEG record for test set.	69
3.9	Examples of IEDs that were missed by the background rejection cascade from 5 EEGs (P1 to P5) for which the algorithm has a low sensitivity.	70
3.10	ROC and precision-recall curves for three classification methods with and without background rejection (BR).	71
3.11	Detection rate per minute versus the number of IEDs in each EEG record for test set (left), and boxplots of detection rates for EEGs with and without IEDs.	71
3.12	False alarm rate (false positives per minute) versus the number of IEDs in each EEG record for test set, along with the box plot for IED-free and epileptic subjects.	73
4.1	Two approaches for IED detection: (a) One single SVM detector (b) SVM cascade followed by one SVM detector.	82

4.2	Sensitivity versus specificity for each subject after a 10-stage classifier cascade.	84
4.3	Precision versus sensitivity at certain specificity values, with and without the classifier cascade.	86
4.4	False positive rate versus sensitivity at certain specificity values, with and without classifier cascade.	86
5.1	Morphological features of an interictal epileptiform discharge. .	96
5.2	The 2-dimensinal scalp EEG electrodes and the X and Y coordinates considered to represent the electrode positions.	101
5.3	Number of scores each expert has given to the events in total. . .	107
5.4	Sensitivity and specificity of experts versus consensus of experts' opinion.	108
5.5	Waveforms with score: 1.5-2.38 sorted from low to high average score. Each waveform lasts 0.6 s at the same amplitude scale. .	109
5.6	Waveforms with score: 2.38-3.3 sorted from low to high average score. Each waveform lasts 0.6 s at the same amplitude scale. . .	110
5.7	Waveforms with score: 3.3-5 sorted from low to high average score. Each waveform lasts 0.6 s at the same amplitude scale. . .	111
5.8	Box plots of morphological characteristics of IEDs for all events, and events scored ≥ 4	112
5.9	t-SNE plot using all features, labelled based on all 18 scorers. .	113
5.10	t-SNE plot using all features, labelled based on top 7 scorers. . .	114
5.11	AC2 agreement coefficient among the groups of randomly selected scorers of size n.	115
5.12	AC2 agreement coefficient among the groups of top scorers of size n.	115
5.13	Feature categories and their maximum correlation with the average and standard deviation of the expert scores.	116

5.14	Scatter plots of some chosen features after feature selection versus expert scores. Pearson correlation values between features and scores are shown on the plots.	120
5.15	Scatter plots of some chosen features after feature selection versus expert scores. Spearman correlation values between features and scores are shown on the plots.	120
5.16	Features selected in feature selection step, which are more significant in predicting the average scores. ψ represents wavelet basis function, ϕ_k is the NLEO with resolution parameter (k).	121
5.17	Features selected in feature selection step which are more significant in predicting standard deviation of the scores. ψ represents wavelet basis function.	122
5.18	The wavelet basis functions listed in Table 5.2 and Table 5.3 which have the highest correlation with the mean and standard deviation of the expert scores.	123
5.19	Root mean square errors resulted from SVR models fitted to the dataset, considering groups of top scorers of different sizes ($n = 2$ to 18); $n = 18$ indicates the group of all experts.	125
5.20	Predicted scores produced by SVR trained using elastic net selected features versus original expert scores, using EEG signals normalized by local standardization	127
5.21	ROC curve for classification based on the SVR output.	128
5.22	Precision-recall curve for classification based on the SVR output.	129
6.1	False positive rate versus sensitivity for the P13 software for three levels of experts agreement [137].	135
6.2	A sample EEG segment considered in the current dataset, displayed and annotated by EEGnet software.	138
6.3	Mean pairwise Gwet's AC2 for all 19 annotators.	143
6.4	Mean pairwise Gwet's AC2 for P13 in thresholds from 0.1 to 1.	144

6.5	FPR versus sensitivity for P13 and human experts for sets of spikes marked by at least 5 experts. P13's performance is shown at the thresholds of 0.1 to 1. P13's and experts' performance is calculated based on the same ground truth and plotted with same color and style for all thresholds.	145
6.6	FPR versus sensitivity for P13 and human experts for sets of spikes marked by at least 8 experts. P13's performance is shown at the thresholds of 0.1 to 1. P13's and experts' performance is calculated based on the same ground truth, and plotted with same color and style for all thresholds.	146
6.7	FPR versus sensitivity for P13 and human experts for sets of spikes marked by at least 12 experts. P13's performance is shown at the thresholds of 0.1 to 1. P13's and experts' performance is calculated based on the same ground truth, and plotted with same color and style for all thresholds.	147
6.8	FPR versus sensitivity for the P13 algorithm and human readers for P13 and human experts for sets of spikes marked by at least 15 experts. P13's performance is shown at the thresholds of 0.1 to 1. P13's and experts' performance is calculated based on the same ground truth, and plotted with same color and style for all thresholds.	148
6.9	Average pairwise performance for P13 at different thresholds and human experts. P13's performance is calculated at the thresholds of 0.1 to 1.	148
6.10	Accelerated bootstrap performance test on the sensitivity and FPR values for P13's perception threshold varied from 0.1 to 1. . . .	149
6.11	Non-inferiority test outcome applied on EEG datasets of size 2 to 200 at 43 PTVs for P13 smoothed by 30 repetitions.	149

6.12	Non-inferiority test outcome for scorer groups of size 3 to 12 averaged over all 43 PTVs, versus the Gwet's AC2 IRA of each group. (b) the average years of EEG fellowship training, (c) the fraction of the group with ABCN certification, and (d) the fraction of the group with ABPN-E certification.	150
6.13	Non-inferiority test outcome for scorer groups of size 3 to 12 averaged over all 43 PTVs, versus the fraction of the group with ABCN certification	150
6.14	Non-inferiority test outcome for scorer groups of size 3 to 12 averaged over all 43 PTVs, versus the fraction of the group with ABPN-E certification.	151
6.15	Non-inferiority test outcome for scorer groups of size 3 to 12 averaged over all 43 PTVs, versus the average years of EEG fellowship training.	151

List of Tables

2.1	Algorithms applied for IED detection, and their performance characteristics reported by authors.	41
2.2	Characteristics of EEG datasets considered in previous studies for IED detection.	44
3.1	Summary of all computed feature categories and their subgroups and parameters.	54
3.2	Training and testing sensitivity for every 5 fold (F1 to F5), as well as the average over all cross-validation folds.	66
3.3	Training and testing specificity for every 5 fold (F1 to F5), as well as the average over all cross-validation folds.	66
3.4	The features selected at every step of the cascade in the training phase.	67
3.5	Overall computation time (in seconds) on the test set for one minute of 19-channel EEG data.	69
4.1	Sensitivity of the classifier cascade after each stage.	83
4.2	Specificity of the classifier cascade after each stage.	83
4.3	Overall performance of IED detection by applying SVM with and without the initial classifier cascade.	85
5.1	Mean values for morphological features of all events, and events scored 4, and above it.	111

5.2	Highest correlation between the features from each category with the mean expert scores. ψ represents wavelet basis function, and ϕ_k is the NLEO with resolution parameter (k). Subscripts ' t ' and ' f ' indicate the operation on PSD values along time or frequency axis, respectively.	117
5.3	Highest correlation between the features from each category with standard deviation of expert scores. ψ represents wavelet basis function, and ϕ_k is the NLEO with resolution parameter (k). Subscripts ' t ' and ' f ' indicate the operation on PSD values along time or frequency axis, respectively.	118
5.4	Correlation coefficient between electrode coordinates, and mean and standard deviation of expert scores.	119
5.5	RMSE values of SVR models based on average expert score for all EEGers for normalized and non-normalized data types, as well as various sets of inputs to the algorithm. The errors are calculated both for predictions made by peak-centered and shifted signals.	126

List of Acronyms

ABCN	American Board of Clinical Neurophysiology
ABPN	American Board of Psychiatry and Neurology
ANN	Artificial Neural Network
AUC	Area Under the Curve
CDF	Cumulative Density Functions
CNN	Convolutional Neural Network
CV	Cross-Validation
CWT	Continuous Wavelet Transform
DAE	Deep AutoEncoder
DBN	Deep Belief Network
DWT	Discrete Wavelet Transform
ECG	Electrocardiogram
EEG	Electroencephalogram
EEGer	Electroencephalographer
EMG	Electromyogram
EOG	Electroculogram
ET	Epileptiform Transient
FP	False Positive
FPR	False Positive Rate
ICA	Independent Component Analysis
IED	Interictal Epileptiform Discharge
ILAE	International League Against Epilepsy

IRA	Inter-Rater Agreement
IRB	Institutional Review Board
k-NEO	k-point NonLinear Energy Operator
kNN	K-Nearest Neighbor
LSTM	Long Short-Term Memory
MGH	Massachusetts General Hospital
MUSC	Medical University of South Carolina
NI	Non-Inferiority
NLEO	Nonlinear Energy Operator
NRMSE	Normalized Root Mean Squared Error
PCA	Principal Component Analysis
PSD	Power Spectral Density
PTV	Perception Threshold Value
PV	Perception Value
RBF	radial basis function
rEEG	Routine scalp Electroencephalogram
RF	Random Forest
RMS	Root Mean Square
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
Se	Sensitivity
SNR	Signal to Noise Ratio
SOFM	Self-Organizing Feature Map
SSE	Summed of Square Error
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
SVR	Support Vector Regression
TM	Template Matching
TP	True Positive

t-SNE t-Distributed Stochastic Neighbor Embedding
WOC Wisdom Of the Crowd

Chapter 1

Introduction

In this chapter, we present the motivation for the research performed in this thesis work, its objectives, and the approach to solve the problem. At the end, we provide a summary of the contributions of this thesis, and the outline of the thesis.

1.1 Motivation

The routine scalp electroencephalogram (EEG) is the most common clinical neurophysiology procedure which includes 20-30 min recording of about twenty electrodes on scalp. Approximately 650,000 outpatient EEGs were performed in the United States in 2013 with a cost of around \$480 million [1].

EEGs of patients with epilepsy might be characterized by occasional interictal epileptiform discharges (IEDs), also called epileptiform transients (ETs), or spikes, occurring between seizures [2, 3]. IEDs can be spikes lasting for 20-70 ms or sharp waves with pointed peaks lasting for 70-200 ms. One sample IED in an EEG recording from Medical University of South Carolina (MUSC) is illustrated in Fig. 1.1.

Presence of IEDs is predictive of recurrence in patients following a first seizure [4, 5]. It is also beneficial in supporting the diagnosis of epilepsy [6].

Figure 1.1: Epileptiform spike (IED) in one EEG recording from MUSC dataset.

Since the presence of IEDs in EEG is often associated with epilepsy, IEDs are instrumental in the diagnosis of epilepsy. However, it is difficult to detect IEDs, because their morphologies are of wide variety, and they can be similar to waves which can be part of the normal background activity and artifacts (such as extracerebral potentials from muscle, eyes, heart, electrodes, etc.). As shown in Fig. 1.2, we can observe an obvious variety in the morphologies of IEDs. These IEDs are extracted from EEG recordings of Massachusetts General Hospital (MGH).

Except for the introduction of digital acquisition, digital display, and digital montage reformatting, not much change has occurred in the way that EEG is recorded and interpreted over the last several years. Currently, an expert detects IEDs in EEGs by visual inspection of 10–20 s of the EEG signals at a time. [7].

In addition, EEGs are frequently misinterpreted by neurologists without neurophysiology fellowship training, who are the majority of neurologists annotating EEGs [8]. Although in current clinical applications, the gold standard for interictal EEG interpretation is still visual inspection and manual annotation, it is tedious and subjective. There is also considerable variability of interpretation between board-certified academic clinical neurophysiologists [7]. Furthermore, experienced EEGers are in short supply [9]. Misinterpreting the EEG may adversely affect patients, as it leads to misdiagnosis of epilepsy and consumption of

Figure 1.2: Examples of different varieties of IED morphology from MGH dataset.

anti-epileptic medications probably for many years; in addition, the treatment of the true underlying cause of the seizure-like events will be delayed. These causes could be cardiac arrhythmias, psychogenic events, or other types [8]. On average, patients with non-epileptic events endure seven years of anti-epileptic medication before being correctly diagnosed [7, 10]. Thus, software for EEG interpretation can be useful to enhance the quality of annotations for neurologists who do not interpret EEG frequently and/or lack neurophysiology fellowship training [3].

Consequently, traditionally experts detect the IEDs from EEG recordings by visual inspection which is tedious and time consuming. Moreover, there is substantial disagreement between experts in EEG interpretation. Hence, automated IED detection is sorely needed, and will increase the uniformity in EEG interpretation of epileptic patients.

There are currently no commonly-accepted IED-detection software programs available [3]. According to few studies performed and according to the commonly-held opinion of academic EEG interpreters, current IED-detection systems have an insufficient sensitivity, which requires an expert to look over

each EEG to make sure that IEDs are not missed [11]. They also have an insufficient specificity, hence, an expert is required to look through a large list of detections which may include many false positives [3, 7, 12, 13].

Moreover, one common problem with several methods that have been applied for automated IED detection is the lack of a sizable database of different IED profiles to validate the performance of the IED detection systems. Therefore, the results are not reliable for clinical purposes. Another problem is that in most studies, the number of interpreters is not large, presumably to minimize study cost [7].

Furthermore, developing accurate automated IED-detection software can be beneficial for enhancing patient care by improving the accuracy of EEG annotation by neurologists with limited training in EEG interpretation. It can also increase the speed and accuracy of annotating prolonged outpatient EEG recordings that last minimum several hours, which will probably become more common in the future as recent studies have shown that EEG recordings longer than the usual 20–25 min lead to improvement in IED detection [14, 15]. IED detection algorithms can also be applied to prolonged inpatient EEG recordings, for instance, in epilepsy monitoring units or intensive care units; these IED detection algorithms need to be trained particularly for inpatient EEG monitoring [7].

To summarize, some important obstacles in IED detection can be listed as:

- High variety of IED morphologies,
- Similarity of IEDs with artifacts as well as normal background EEG activity,
- Low degree of agreement among experts in IED recognition,
- Not enough number of experts who annotate EEG datasets ,
- Lack of sufficiently large annotated EEG dataset of patients with epilepsy.

Due to the current short-comings of IED detection methods, and the significance of automated IED detection for diagnosis and management of epilepsy,

there is currently a high demand for automated, accurate, and quick IED detection systems which can be reliably applied for clinical purposes.

1.2 Objectives

The goal of this study is to discover underlying patterns in IEDs and expert interpretation of IEDs, and to develop a fast IED detection system which can accurately detect IEDs in the EEG of patients with epilepsy. Large datasets of EEG annotated by expert neurologists are utilized in order to develop and validate the methods and algorithms proposed in this research work. We need to train the algorithms using very large datasets in order to consider almost all possible IED morphologies with the purpose of making reliable detections. Nevertheless, speed is one substantial issue in IED detection methods based on large datasets. Therefore, we need to consider algorithm speed as a non-negligible factor.

This thesis consists of different parts, each of which aims to contribute to automated IED detection in a certain way. For each aspect, a specific scalp EEG dataset is employed which best suits the purpose of the study. In this thesis, algorithms are developed for IED detection. In addition, the agreement in IED detection among the experts is analyzed in order to learn more about the aspects of IEDs which lead to consistent detection. Moreover, methods are proposed to determine whether an algorithm can perform similar to a human expert or not, and an IED detection software is tested as an example. Each part of this research is described as follows:

I. Background rejection algorithm: A large amount of data is needed for training and evaluating the performance of an effective IED detection system. IEDs occur infrequently in the EEG of most patients, hence, interictal EEG recordings contain mostly background waveforms. In order to be able to cope with this vast volume of data, we propose a feature-based ensemble method in

this work. This method is a similar approach employed by Viola *et al.* [16] for the problem of face detection. We develop a system of simple classifiers to eliminate most of the background waveforms in the EEGs. This algorithm reduces the amount of data and helps in speeding up IED detection.

Therefore, as the first step to detect IEDs automatically, we propose to eliminate most EEG background data using simple fast classifiers. This could save computation time for an IED detection system, allowing the remaining waveforms to be classified by more computationally intensive methods. We design a feature-based cascade and employ it to eliminate background data in interictal EEG while preserving most IEDs.

We analyze a dataset of EEG recordings from 156 subjects from Massachusetts General Hospital (MGH). Since IED characteristics are of great variety, the proposed method consists of multiple stages to reject the background waveforms, with each stage employing one specific feature. We choose several quick-to-compute EEG features. These features include: (1) morphological features, (2) nonlinear energy operator (NLEO), (3) discrete and continuous wavelet transform coefficients, (4) energy of the signal, (5) features in main EEG frequency bands, and (6) spectrogram and power spectral density.

We calculate empirical cumulative density functions (CDF) for each of these features for IED and background waveforms separately. We rank the features based on them and select one feature in each stage. By thresholding these features in consecutive steps, background waveforms are rejected sequentially. We then apply three classifiers to detect IEDs with and without initial background rejection, with the purpose of speeding up the classification.

II. Cascade of classifiers: Next, we propose another ensemble method comprising a cascade of support vector machines (SVMs) to classify EEG data using the raw EEG signal. In this method, a huge number of SVMs compete in each step of the cascade in terms of their specificity values at a fixed sensitivity. One

part of classification task is done in each stage by the winner classifier. The difference between classifiers is the set of waveforms they have been trained by, i.e., we train each of them by a different set of background waveforms.

As mentioned before, current IED detection methods suffer from insufficient precision and high false positive rates. This method is proposed with the goal of improving the precision and false positive rate by incorporating the classifier cascade with IED detection.

III. Inter-rater agreement analysis: Using a set of IEDs scored by a large number of experts, we set out to determine which attributes of IED predict expert agreement regarding the presence of IED. We consider EEG waveforms from 200 patients annotated by a large number of neurologists provided by Medical University of South Carolina (MUSC). IEDs were annotated on a 5-point scale by 18 clinical neurophysiologists within 200 30-second EEG segments from recordings of 200 patients. We extract numerous signal analysis features to represent the data. We apply feature selection and predict expert agreement. We also investigate different normalization schemes. We try to improve IED detection by understanding more about the features which optimally characterize IEDs, and the features which lead to agreement or disagreement among experts in IED recognition.

IV. Comparison of software with human experts: Last but not the least, we develop methods to evaluate the performance of any IED detection algorithm. These measures are applied on Persyst 13 (P13) algorithm which has recently claimed to have passed a “statistical Turing test” for detecting IEDs, achieving algorithm-human inter-rater agreement (IRA) similar to human-human IRA at a perception threshold value (PTV) of 0.1. To further investigate, we compare P13’s performance on an EEG dataset from MUSC containing 200 patients with the performance of a large number of expert physician electroencephalographers (EEGers).

1.3 Major Contributions of the Thesis

We list the contributions of this thesis work for the problem of automated IED detection below.

We designed a cascade of thresholding steps and employed it to eliminate most background data in the interictal EEG while preserving majority of IEDs. Background waveforms were rejected by applying thresholds on the features in consecutive steps. We selected features based on their empirical cumulative density functions (CDF). We applied the proposed algorithm on a dataset consisting of 156 epileptic and IED-free subjects. The algorithm was able to eliminate 78.41% of background waveforms while retaining 96.93% of IEDs on our cross-validated dataset. Furthermore, we applied classifiers to detect IEDs with and without initial background rejection. The results show that rejecting background by our proposed method speeds up the classification by a factor ranging from 4.69 to 1.76 for the considered classifiers.

In the next step, we designed an ensemble method comprising a number of support vector machines (SVMs) in the form of a cascade. We trained each SVM using a different set of background waveforms. These classifiers compete in each stage, and one SVM is chosen each time to classify the waveforms. We compared the IED detection performance of a single classifier with the same classifier following the proposed cascade. Results show that the precision and false positive rate improve significantly by incorporating the classifier cascade with an IED detection method. This method can help enhance the performance in terms of precision and false positive rate of an IED detection system. At a fixed sensitivity, we were able to improve precision by 6.78%; and at a fixed false positive rate, the sensitivity improved by 2.83%.

We performed a thorough analysis on expert inter-rater agreement in annotating IEDs and features which correlate with this agreement. We visualized the waveforms in 2D by the t-SNE algorithm using the extracted features which

provide us with good separation between low and high scored events.

Using feature selection and correlation analysis, we determined subsets of features which correlate with expert scoring and contribute to agreement or disagreement among experts. We fitted regression models in order to predict scores given by EEGers to each event based on the extracted features. We investigated the effect of normalization on interictal EEG analysis as well. In addition, we analyzed the relation of electrode locations on the scalp with expert scores and determined the locations for which experts give higher scores.

We found multiple feature types to be useful for predicting expert annotations, however, particular types of wavelet features performed best. Additionally, local normalization improved predictability, suggesting the importance of IED morphology over amplitude-based features. As the size of the group of EEGers considered to train the models was increased, the performance of the models levelled off at a group size of around 11. Although most IED detection studies in the past have used opinion from three or fewer experts, our study suggests a “wisdom of the crowd” effect, such that pooling over a larger number of expert opinions produces a better correlation between expert opinion and objectively quantifiable features of the EEG.

We developed and implemented a group of statistical tests for evaluating the reliability of an IED detection system, and to investigate whether it is non-inferior to human experts. We considered P13 algorithm as a case study. We applied methods such as pairwise ROC analysis, pairwise inter-rater agreement analysis. We also applied accelerated bootstrap analysis in different modalities to determine non-inferiority of the algorithm. We also investigated the effect of expert characteristics by this method. According to the results, this algorithm fails to perform as well as clinical neurophysiologists based on mean pairwise sensitivity and false positive rate and performs best at the perception threshold value setting of greater than 0.8 on our dataset. Therefore, it was found out that P13 is not as reliable as a human expert in detecting IEDs on our dataset.

No commercially available software has yet convincingly passed the “statistical Turing test” for IED detection. Therefore, the problem of IED detection still remains unsolved.

1.4 Organization of the Thesis

The remaining chapters of this thesis are organized as follows: In Chapter 2, we present the literature survey and background of IED detection research. Chapter 3 includes the proposed background rejection algorithm for speeding up IED detection. Next, we present the proposed algorithm using cascade of SVMs to improve the false positive rate and precision in Chapter 4. Chapter 5 includes inter-rater agreement analysis methods to find patterns and features underlying expert agreement. In Chapter 6, we compare human experts with an IED detection software and explain the proposed methods to test its performance. Chapter 7 contains the conclusions from the research performed in this thesis work and suggestions for the future work.

Chapter 2

Literature Survey

In this chapter, we provide a review on the background of spike detection research and the main methods that have been applied in the literature for this purpose; and we compare these methods. In addition, we explain the studies on expert agreement analysis in spike detection performed in previous studies.

2.1 Definition of Terms

In this section, we provide the definitions and background for the concepts and terms related to epilepsy diagnosis.

2.1.1 Epilepsy

Epilepsy is a set of neurological diseases marked by epileptic seizures [17] which is also known as seizure disorder. Seizures are the most obvious visible symptom of epilepsy, lasting from a few seconds to a few minutes. Seizures may involve loss of consciousness, atypical or odd behavior, distorted perceptions and typically happen without warning. Patients with seizures attack may fall to the ground with muscles stiffen or jerk out of control.

Epilepsy can affect people of all ages. It is one of the most common neu-

rological diseases. Approximately 50 million people worldwide have epilepsy; and about 80% of the patients with epilepsy live in low and middle income countries [18].

The definite cause of epilepsy remains unknown for most cases. However, there are certain factors that can be given: some people suffer from epilepsy as the result of brain injuries, brain infections (meningitis, encephalitis), cerebral palsy, stroke due to arterial blockage (Hemorrhage), brain tumors; some patients are infants with abnormal brain structures; some develop epilepsy with other systemic diseases. Genetic mutations and drug abuse are also directly linked to a small ratio of epilepsy cases [19]. Although the basic cause of epilepsy is complicated and diversified, the direct cause of seizures is disturbances in the electrical activity of the brain, i.e., exceptional and abnormal nerve cell activity in the brain cortex [20].

According to Epilepsy Foundation [21], the epileptic seizures may be related to a brain injury or a family tendency, but the cause is often unknown. Many patients experience more than one type of seizure and can also have other symptoms of neurological conditions. In the situations that EEG testing, clinical history, family history, and outlook are similar for a group of patients, their condition could be defined as a specific epilepsy syndrome.

Epilepsy has been defined conceptually in 2005 as "a disorder of the brain characterized by an enduring predisposition to generate epileptic seizures." This definition is usually applied as "having two unprovoked seizures >24 h apart." A task force altered the practical definition for special circumstances that do not meet the two unprovoked seizures criteria; the International League Against Epilepsy (ILAE) [22] accepted their recommendations. The revised definition of epilepsy brings the term in concordance with common use. It was proposed that epilepsy would be considered to be "a disease of the brain defined by any of the following conditions: (1) At least two unprovoked (or reflex) seizures occurring >24 h apart; (2) one unprovoked (or reflex) seizure and a probab-

ity of further seizures similar to the general recurrence risk (at least 60%) after two unprovoked seizures, occurring over the next 10 years; (3) diagnosis of an epilepsy syndrome." Epilepsy is considered to be "resolved" for people who "either had an age-dependent epilepsy syndrome but are now past the applicable age, or those who have remained seizure-free for the last 10 years and off anti-seizure medicines for at least the last 5 years" [17].

2.1.2 Electroencephalography

Electroencephalography (EEG) is an electrophysiological monitoring technique recording electrical activity of the brain which might be useful in diagnosing brain disorders, especially epilepsy or another seizure disorder. Scalp EEG is noninvasive, as the electrodes are placed on the scalp; while invasive electrodes can also be used such as in electrocorticography [23]. Diagnostic applications typically focus on event-related potentials to investigate potential fluctuations time locked to an event like stimulus onset or button press; or they focus on the spectral content of EEG to analyze the type of neural oscillations observed in the EEG in the frequency domain.

EEG is utilized in the diagnosis for a number of disorders; and the total number of EEGs performed each year is not a reportable event. However, it is estimated to be in the range of 10–25 million [24].

EEG is most often used to diagnose epilepsy [25]. It is also exploited to diagnose brain damage from head injury, brain tumor, brain dysfunction which may have different causes (encephalopathy), inflammation of the brain (encephalitis), stroke, sleep disorders, depth of anesthesia, coma, and encephalopathies. It can also be used to confirm brain death for a person in a persistent coma. Moreover, it is utilized to find the right level of anesthesia for a person in a medically induced coma [25]. EEG continues to be a valuable tool for research and diagnosis despite its limited spatial resolution. Moreover, EEG is one of the few available mobile techniques, and provides temporal resolution in the range of

millisecond which is not possible with CT, PET or MRI [26].

2.1.3 Interictal epileptiform discharges (Spikes)

Epileptiform spikes are biomarkers for the basic pathophysiology of the epileptic condition. In most clinical practice, epileptiform spikes in EEG recordings are utilized to help confirm a diagnosis of epilepsy and identify a patient's epilepsy syndrome. Interictal spikes are distinctive epileptiform discharges appearing between seizures and can be viewed as prognostic of seizure attack. According to the guidelines of International Federation of Societies for Electroencephalography and Clinical Neurophysiology (IFSECN) [27], there are different kinds of interictal spikewaves: sharp wave, spike, slow wave, spike-and-slow-wave complex, multiple spike complex, polyspike-and-slow-wave complex, etc. The primary terms are defined as follows.

'Epileptiform pattern' which is synonym with 'epileptiform discharge' and 'epileptiform activity' describes "transients distinguishable from background activity, with a characteristic spiky morphology, typically, but neither exclusively nor invariably, found in interictal EEGs of people with epilepsy" [28].

A 'transient' in EEG is defined as "any isolated wave or complex (a sequence of two or more waves having a characteristic form or recurring with a fairly consistent form, distinguished from background activity.), distinguished from background activity" [28]. There are several types of waveforms which can be interpreted as epileptiform patterns, such as

Spike: "A transient, clearly distinguished from background activity, with pointed peak at a conventional paper speed or time scale and a duration from 20 to under 70 ms, i.e. $1/50 \pm 1/15s$, approximately. Main component is generally negative relative to other areas. Amplitude varies. This term should be restricted to epileptiform discharges. EEG spikes should be differentiated from sharp waves, i.e. transients having similar characteristics but longer durations." However, this distinction is largely arbitrary and primarily serves descriptive pur-

poses. Generally, in ink-written EEG records taken at 3 cm/s, spikes occupy 2 mm or less of paper width and sharp waves more than 2 mm.

Also, EEG spikes should be clearly distinguished from the brief unit spikes recorded from single cells with microelectrode techniques [28].

Sharp wave: "A transient, clearly distinguished from background activity, with pointed peak at a conventional paper speed or time scale and duration of $70 \pm 200ms$, i.e. over $1/4 \pm 1/5s$ approximately. Main component is generally negative relative to other areas. Amplitude varies." This term should be restricted to epileptiform discharges and does not apply to (a) distinctive physiologic events such as vertex sharp transients, lambda waves and positive occipital sharp transients of sleep, (b) sharp transients poorly distinguished from background activity and sharp-appearing individual waves of EEG rhythms."

Also, Sharp waves should be differentiated from spikes, i.e. transients having similar characteristics but shorter duration. However, this distinction is largely arbitrary and primarily serves descriptive purposes. As a rule, in ink-written EEG records taken at 3 cm/s, sharp waves occupy more than 2 mm of paper width and spikes 2 mm or less [28].

Spike-and-slow-wave complex: "A pattern consisting of a spike followed by a slow wave [28]."

Polyspike-and-slow-wave complex: "Sequence of two or more spikes associated with one or more slow waves"

polyspike complex: "A sequence of two or more spikes" [28].

They are all collectively named "interictal epileptiform discharges (**IED**)" or "**spike**" in this dissertation for consistency.

2.2 IED Detection Methods

In this section, we explain the algorithms applied for IED detection in the literature. Good reviews of spike analysis studies can be found in [3, 29].

The main methods which have been applied for automated IED detection include template matching, parametric or autoregressive methods, Mimetic analysis, power spectral analysis, wavelet analysis, and artificial neural networks. Nevertheless, many approaches benefit a combination of the methods; some apply artifact rejection techniques [3]; and several recent algorithms have employed newer techniques. These methods are described in the remainder of this section.

2.2.1 Template matching

Template Matching (TM) is the process of detecting an event buried in a signal by comparing it to a predefined template. The goal is to locate possible events in the signal that closely resemble the template. The distance measure definitions employed widely by the template matching algorithm are Summed of Square Error (SSE), convolution, and cross-correlation [30]. In practice, the basic method is to determine how close any given section of a signal is to the template. A threshold is applied on the distance or similarity measure, in order to find the events with highest similarity with the template.

For the purpose of IED detection, a model IED waveform is chosen as a template, and whenever the cross-correlation, or other similarity measure, of the EEG data with the template exceeds a threshold, an IED is detected. Stevens *et al.* [31] made the first attempt using this method for IED detection. This method has also been applied in [32]. Lodder *et al.* [33] developed a self-adapting TM system based on a set of smart templates. The study was conducted on 23 patients with 723 IEDs. The sensitivity was reported as 95% and false positive per minute as 6.6, using 15 iterations. Nonclercq *et al.* [34] applied temporal clustering to on 2500 IEDs from 3 patients, and created a template library. A system which adapts to inter-patient and intra-patient variation in spike morphology was developed based on the template library. Sensitivity and selectivity were reported as 93.7% and 93.7%, respectively.

Two measures commonly applied in TM for IED detection, are briefly

described below:

- TM can be carried out based on the z-normalized Euclidean distance, which is equivalent to SSE, as in [35]. Euclidean distance between two n -dimensional samples, p and q , is denoted by $d(p, q)$, and is calculated as:

$$d^2(p, q) = \sum_{i=1}^n (p_i - q_i)^2. \quad (2.1)$$

- In [34], cross-correlation between all signal frames and the template is employed to measure the similarity between template and EEG segments. Cross correlation between two time series x_t and y_t , is computed as [36]:

$$\rho_{xy}(m) = \frac{(x_n - \mu_x)(y_{n+m} - \mu_y)}{\sigma_x \sigma_y}, \quad (2.2)$$

where, μ_x and σ_x denote mean and standard deviation of x_n , respectively. E indicates expected value (mean).

The main factor in TM is choosing the template, since the output is highly dependent on the matching template. The problem of choosing appropriate templates which are sufficient to match with almost all possible IEDs, can limit the performance of this method.

2.2.2 Parametric or autoregressive models

We may model EEGG signals using a mathematical model which is characterized by a set of parameters. If the mean, variance, and autocorrelation function of a signal do not change over time, it is considered stationary. In the parametric methods, the background EEG activity is considered stationary. Whenever a difference is detected between the waveform which the model predicts and the recorded signal, it indicates the occurrence of a non-stationary part in the EEG, and an IED is detected [32, 37–42].

In the study by Sankar and Natour [41] EEG signal was modeled as a

weighted linear combination of its past values and the present value of a hypothetical input, u_n to a filter whose output is the EEG signal, s_n . Equivalently, in the frequency domain, EEG was modeled as the output of an all-pole filter excited by the hypothetical input. EEG signal is represented as:

$$S_n = - \sum_{k=1}^p a_k S_{n-k} + G u_n, \quad (2.3)$$

where, the model parameters a_k are the linear prediction coefficients, p is the model or prediction order, and G is the gain. Signal s , is approximated from the linearly weighted summation of its past samples, which is denoted by $\hat{S}_n$:

$$\hat{S}_n = - \sum_{k=1}^p a_k S_{n-k}, \quad (2.4)$$

and the residual error is:

$$e_n = S_n - \hat{S}_n = G u_n. \quad (2.5)$$

The a_k parameters are derived by minimizing the total squared error with respect to each of the parameters, by using least square analysis. It results in a set of equations called normal equations:

$$\sum_{k=1}^p a_k \sum_{n=0}^{N-1-i} S_{n-k} S_{n-i} = \sum_{n=0}^{N-1-i} S_n S_{n-i}, \quad (2.6)$$

utilizing the autocorrelation, we have:

$$\sum_{k=1}^p a_k R(i-k) = -R(i), \text{ for } 1 \leq i \leq p \quad (2.7)$$

where the autocorrelation function is:

$$R(i) = \sum_{n=0}^{N-1-i} S_n S_{n+i}, \text{ for } i \geq 0 \quad (2.8)$$

where, N is the number of samples in the signal window. The predictor coefficients a_k , can be obtained by solving the normal equations.

The LPC model parameters were calculated by a fast, recursive Levinson-Durbin algorithm [43]. The inverse filter $A(z)$ is determined by the model parameters:

$$A(z) = 1 + \sum_{k=1}^p a_k z^{-k}, \quad (2.9)$$

and

$$H(z) = \frac{G}{A(z)} \quad (2.10)$$

where $H(z)$ is an system's all-pole transfer function. EEG can be processed by the inverse filter which produces the residual as in Fig. 2.1. After determining the residual function, its autocorrelation is derived for sharpening the non-stationarities of EEG.

The algorithm searches for the residual, for a certain number of consecutive samples having amplitudes greater than a specific value. If the above conditions are met, an IED is detected.

The usefulness of parametric methods is limited, since the IEDs turn out not to be less stationary than other types of EEG signals such as artifacts. In addition, these methods are very sensitive to the number of terms employed to estimate the signal, and if we don't have some a priori knowledge to select the number of terms, we would have fragile models [29].

2.2.3 Mimetic analysis

Mimetic methods analyze the signal shape in a similar way as humans describe them. The EEG waveform is decomposed into segments between the amplitude extrema, each of which is called a half-wave. These two half-waves with opposite directions, form the triangular shape of an IED. Having some predetermined criteria for duration, amplitude, slope, sharpness, emergence from background,

Figure 2.1: Linear prediction model used by Sankar and Natour [41].

etc., we can analyze the half-waves if they fit these criteria. This method has been applied in many studies including [44–51].

Periklis *et al.* [52] performed one of the first studies on interictal spike detection. They applied a digital computer technique by which abnormal scalp EEG transients were magnified and processed so that detailed characteristics of their duration and sharpness could be precisely quantified. They digitized and stored the spikes on computer-compatible magnetic tape. The digital data was then low-passed and differentiated by digital filters. And they offered the option of acquiring a picture of the processed data via an interface with a storage and display unit. They quantified the duration and sharpness characteristics of spikes. They proposed a set of six parameters describing these characteristics and investigated spikes chosen from three epileptic EEGs. They concluded that the proposed method was superior to quantifying abnormal EEG transient characteristics directly from polygraph paper.

One example of mimetic parameters applied in IED detection, can be found in the study by James *et al.* [53]. They implemented a combination of mimetic, Artificial Neural Network (ANN), and fuzzy logic. In their study, first a peak is detected by a simple peak detector, and then the parameters are extracted

from the corresponding segment. The parameters include voltages, slopes, and durations, and are shown in Fig. 2.2, and are explained below: Since IED must be

Figure 2.2: Parameters extracted from mimetic analysis [53].

clearly distinguishable from background EEG, contextual parameters were also extracted from the 1.0 s segment of EEG with the peak of each waveform in the center. After computing the parameters, thresholds were set on each of them, and once a waveform was found that exceeds the thresholds, its peak was considered the primary peak, and the detection was made based on that.

The problem with mimetic methods is that, the traditional morphologic definition of IED is not sufficient for IED detection. Since many other transients such as eye blinks, vertex waves, electrode artefact, and movement artefacts, which can be normal, abnormal, or artifactual, might fit the same definition. These transients are common in long EEG recordings.

Some attempts to improve this method include taking first and second derivatives using Walsh operator to determine sharpness of the transients [54, 55], and using Fourier transform in addition to mimetic analysis [56], analyzing additional signals like electroculogram (EOG), electrocardiogram (ECG), and electromyogram (EMG) [57], considering spatial context of EEG by analyzing adjacent EEG channels [44, 56–58], state dependent IED detection using different algorithms depending on the subject’s state (awake, asleep, drowsy) in order to adjust to the occurrence of specific electrocortical or artifactual waveforms in different brain

states [59, 60].

In the analysis by Liu *et al.* [61], all possible IED candidates were detected by using k-point nonlinear energy operator (k-NLEO). A spike model with slow-wave features was then applied to the candidates for IED detection. Adaboost classifier was applied to classify the waveforms. Within their 126 trials, totally 42 single spikes and 100 spikes with slow waves were identified by the experts. For a candidate point P, detected by k-NLEO, four other feature points were defined. Points A, P and B represent the spike, and the points B, Q and R represent the slow wave. Fig. 2.3 shows the feature points used for mimetic feature extraction in this model.

After the five feature points were selected, 13 features were calculated to be used for classification. The 13 features were divided into four groups, including duration, amplitude, slope and area. The conventional spike model for spike detection in Acir's study [62] was formed by six features: Dur_{AP} , Dur_{PB} , Amp_{AP} , Amp_{PB} , $Slope_{AP}$ and $Slope_{PB}$. This feature set was denoted as FS1. Their proposed spike model was designed to include the conventional spike features as well as the slow wave related features ($Dur_{slowwave}$, $Amp_{slowwave}$ and $Area_{slowwave}$). These features were collectively denoted as FS2. FS3 contains the nine features in FS2 and 4 additional spike related features of Dur_{spike} , Amp_{spike} , $Slope_{sharpness}$ and $Area_{spike}$. All feature sets were then used in the classification module to compare spike detection performance. Their results showed that the by adding the slow wave model, accuracy of IED detection could be improved by 6.5% and 21.5% for two and three class classification.

2.2.4 Power spectral analysis

In this method the amount of power in special frequency bands is quantified. Power spectral analysis is a classical method of defining an EEG signal in terms of frequency power quantified in different frequency bands. The most common

Figure 2.3: Feature points and reference diagram for feature extraction [61].

type of this method for IED detection is using Fourier transform [49, 51, 56, 63].

Hilbert transform [64], and Walsh transform [54] have also been applied.

Discrete Fourier transform of signal $x(n)$ is defined as [65]:

$$X(k+1) = \sum_{n=0}^{N-1} x(n+1)W_N^{kn}, \quad (2.11)$$

where,

$$W_N = e^{-j2\pi N}. \quad (2.12)$$

Power Spectral Density (PSD) is defined as [66]:

$$S_{xx}(k) = \lim_{n \rightarrow \infty} E[X^2(k)], \quad (2.13)$$

where $E[\]$ indicates expected value (mean). Equation (2.13) means that we estimate the PSD of a signal by taking expected value of magnitude squared of its Fourier transform.

For instance, in the study by Exarchos *et al.* [51], the PSD of each event was divided into 10 distinct frequency ranges, and PSD for each range was derived as a feature; hence, 10 features were extracted for each event.

A limitation with these methods is that they consider a fixed time-frequency resolution. Increasing the resolution in either time or frequency decreased the resolution in the other. Therefore, this method has largely been replaced by methods like Wavelet analysis.

2.2.5 Wavelet analysis

The wavelet transform employs a multi-resolution approach which represents higher-frequency components with higher time resolution. This method is beneficial in representing different aspects of EEG signals, for instance, discontinuities, and repeated patterns where other signal processing approaches fail. By using wavelet transform, we can characterize parts of EEG signal, which contain non-stationary transient events. Therefore, it is suitable for IED analysis. Many recent IED detection algorithms have applied this method [13, 42, 67, 68].

Wavelet Transform is a powerful tool for signal analysis similar to Fourier transform. Fourier transform decomposes the signal into a series of sine waves with different frequencies, while wavelet transform decomposes the signal into scaled and shifted versions of the mother wavelet. It hence allows localization in both time and frequency domains by means of translation and scaling. There are two types of wavelet transform, i.e., continuous and discrete wavelet transforms; these techniques are described as follows.

We can exploit orthogonal wavelets for discrete wavelet transform, and non-orthogonal wavelets for continuous wavelet transform. These two transforms have the following properties:

Continuous Wavelet Transform (CWT) applies inner products to quantify the similarity between a signal and an analyzing function which is a wavelet, ψ . CWT compares the signal to shifted, compressed or stretched versions of the wavelet. The CWT coefficients $C(u, s)$, are obtained by varying the values of the scale parameter, u , and the position parameter, s , as in Equation (2.14). In addition, depending on what signal features we are trying to detect, we can select

a wavelet that facilitates the detection of that feature. The CWT of a signal, $f(t)$ using mother wavelet, $\psi(t)$ is given as:

$$C(u, s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} f(t) \psi\left(\frac{t-u}{s}\right) dt, \quad (2.14)$$

where s and u represent the scaling and translation parameters, respectively. The wavelet coefficient C for $\psi_{u,s}(x)$ defines shifting and scaling of the mother wavelet:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right), \quad (2.15)$$

where compression or stretching is determined by the scale parameter s [69].

The Discrete Fourier Transform (DFT) decomposes a signal into sinusoidal basis functions with different frequencies. Information is not lost in the transformation, and the original signal can be recovered from its DFT representation.

Discrete Wavelet Transform (DWT) decomposes the signal into a set of mutually orthogonal wavelet basis functions which are different from sinusoidal basis functions as they are spatially localized, i.e., they are non-zero over only part of the signal length. Wavelet functions are dilated, translated and scaled versions of the mother wavelet ϕ . DWT is also invertible, and the original signal can be completely recovered from its DWT representation [70].

In the methods which employ wavelets for IED detection, the wavelet transform coefficients are derived, and detection is made using these coefficients as features.

Moreover, some previous works, suggest certain wavelet basis functions to be helpful in IED detection. Some main conclusions from previous methods based on wavelet transform are explained as follows:

- In the study by Kalayci and Ozdamar [42], the overall highest classification accuracy, was obtained using Daub-20 wavelet function and the combination of Scales 3 and 4, which correspond to 3.72-8.84, and 1.56-3.72 Hz, respectively. Lower scales that contain higher frequency components do not carry much in-

formation to distinguish IEDs. It could be because high frequency components of IED events and EMG activity may accumulate in the same scales.

- Guler and Ubeyli [68] investigated few wavelets including db2, Coiflet of order 4 (coif4), Symmlet of order 10 (sym10), Daubechies of order 1 (db1), and Daubechies of order 6 (db6). They observed that the Daubechies wavelet yielded better accuracy, and db2 was marginally better than db1 and db6. Thus, they chose db2 for their study.
- Indiradevi *et al.* [13] applied a different approach to choose the wavelet. They computed the cross-correlation between a known IED and different wavelet bases and chose Daubechies 4 to obtain highest correlation coefficients. They focused on sub-bands 4 and 5 (4–8 and 8–16 Hz) in order to minimize contamination by non-spike high frequency signals like myogenic artefacts. Next, they extracted IED portion of EEG from the raw data and determined the signal energy for locating the location of epileptic foci.

As we see from previous studies, the choice of optimum wavelet basis functions and scales, is still an open topic in IED detection research.

2.2.6 Artificial neural networks

Artificial Neural Networks (ANNs) are modeled using properties of natural neurons. They include basic computing units like neurons and are designed to learn and recognize patterns similar to human brain. ANNs approximate generally unknown functions that may depend on a large number of inputs. They comprise interconnected neurons that exchange messages between each other. The connections between neurons have tunable weights, which make ANNs capable of learning.

Most of the ANN methods applied for IED detection employ supervised learning, which means that the training signals are classified before inputting to

the network. However, using self-organizing neural networks, classification of each individual signal is not required, and it is sufficient to just classify the whole EEG recording for training, as in [53].

There is no optimal input for ANNs, several different inputs have been employed including raw EEG signal [71–74], parameters extracted from mimetic analysis [53, 62, 75–78], power spectra produced from a parametric model [79], power spectra from wavelet analysis [40, 68, 80–82]. In some of these methods, expert systems are employed, which operate on the output of ANNs [40, 57, 79, 83, 84]. An expert system is a kind of artificial intelligence algorithm in which a knowledge base is made using accumulated experience and a set of rules, to be applied in different situations.

Results of some previous studies which applied ANNs to detect IEDs, are described below.

In the study by Kalayci and Ozdamar [80], the purpose was to investigate the feasibility of using wavelet transform as a pre-processor for an IED detection system based on ANNs. They attempted to decrease the input size, without decreasing the information content of the signal and degrading the performance. Their results show that applying the wavelet transform drastically decreases the input size of the ANN, without much compromise in its performance. The most important factor in performance was the proper selection of the scale or the combination of scales.

In another study, Ozdamar and Kalayci [72] utilize raw EEG data as the input to a feed-forward ANN with a variable number of input and hidden layer units. The effects of different EEG time windows and the number of neurons in hidden layer were analyzed for highest sensitivity and selectivity. They achieved their best ANN configuration with an input time window of 150 msec (30 input units) and 6 hidden layer neurons. They concluded that raw EEG can be successfully employed for training ANNs for IED detection.

In addition, the performance of their method was improved by adding another

output for EMG detection. The reason for adding EMG detection was that EMG noise can sometimes mimic an IED waveform; therefore, it generates false positive detections. Thus, detection of EMG helps reducing the false positive rate.

In the study by James *et al.*[53], a Kohonen's self-organizing feature map (SOFM) is employed. It is a lattice type ANN, and its neurons are trained using an unsupervised learning process. The structure of this square 2-dimensional SOFM is shown in Fig. 2.4. All neurons are connected to the input, x , via weight vectors m_i .

Synaptic weights of the SOFM are changed through a learning process known

Figure 2.4: The two-dimensional array of neurons in the self-organizing feature maps (SOFMs).[53].

as Kohonen's learning rule which is described below: In Fig. 2.4, all the inputs are connected in parallel to all neurons, i in the network, and each neuron has a corresponding weight vector m_i . The weight vectors are initially assigned to small random values. When an input x is presented, the Euclidean distance is calculated, and the winning weight vector m_c is selected such that:

$$|x - m_c| = \min |x - m_i|. \quad (2.16)$$

Next, a neighbourhood set N_c around neuron c is defined, and the weights of

neurons of that neighbourhood are updated:

$$\begin{aligned} m_i(t+1) &= m_i(t) + \alpha(t)[x(t) - m_i(t)], \text{ if } i \in N_c(t) \\ m_i(t+1) &= m_i(t), \text{ if } i \notin N_c(t) \end{aligned} \quad (2.17)$$

Learning rate $\alpha(t)$ and the neighbourhood function gradually decrease by time:

$$\alpha(t) = \alpha_0 e^{\left(\frac{-t}{\tau_\alpha}\right)}, \quad (2.18)$$

and

$$N(t) = N_0 e^{\left(\frac{-t}{\tau_N}\right)}. \quad (2.19)$$

Next, input x is fed to the network and the procedure is repeated. The weight vectors represent the statistics underlying the input vectors. They form topographic maps of the inputs across the 2D lattice in which similar inputs excite neurons in similar SOFM regions.

In the study by Webber *et al.* [76], a 3-layer feed-forward ANN is employed to detect IEDs using both raw data, and parameterized data from mimetic analysis. According to their results, parameterized data gives slightly better performance. However, they stated that they need to further investigate to come to a conclusion that which data type is better.

In the study by Acir *et al.* [62] a three-stage procedure based on ANNs is proposed. In the first stage, a pre-classification unit is trained by the discrete perceptron learning rule and separates possible non-IEDs and IEDs from definite IEDs and non-IEDs. This reduces the computation time, as well as increasing the overall detection performance of the system. The second stage aims to separate IEDs and non-IEDs by an ANN functioning as a non-linear post-classifier.

Three different ANNs, i.e., a back propagation multi-layer perceptron trained by a hybrid method, and two radial basis function (RBF) networks trained by a support vector method were employed as the post-classifier. They were compared based on their classification performance. In the third stage, multichannel information

was integrated for detecting IEDs. The architecture of the radial basis function network (RBFN) used by Acir *et al.* [62] is shown in 2.5. The best performance was obtained with the RB-SVM.

In addition, they stated that, the detection of IEDs using raw EEG data, may have the potential advantage of avoiding the possible false classification which may occur due to loss of data in the EEG parameterization. To combine the advantages of two methods, they have utilized both the parameterized and raw input data in the pre-classification and post-classification stages, respectively.

Figure 2.5: Architecture of the radial basis function network (RBFN) used by Acir *et al.* [62].

Carey *et al.* [85] used a simple ANN design utilizing only windowed voltage data from only one single EEG channel. Autocorrelation of a template IED was computed to provide a baseline comparison. Their method achieved mean precision of 72.67% and sensitivity of 82.68% for 6 test patients, while the autocorrelation mean precision and sensitivity were 60.43% and 90.61%, respectively. Their ANN design employed the standard feed forward algorithm using the back propagation algorithm. It contained a single hidden layer with 10 neurons. The input layer had 120 neurons; hence, the input was 120 values from the windowed datasets. The architecture of this ANN is shown in Fig. 2.6. According to the

Figure 2.6: The feed forward artificial neural network (ANN) using error back propagation, by Carey *et al.* [85].

studies that employ ANNs for IED detection, we see that there is no optimum input type which authors agree on. In addition, there are several types of ANNs which can be applied for this purpose, and the choice is almost flexible.

2.2.7 Deep learning

Two most common deep learning architectures are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs are motivated by Hubel and Wiesel's work on the cat visual cortex [86] and they process input data as multidimensional arrays, preserving spatial information. They usually comprise multiple convolutional and pooling layers in addition to their input layer (receptive field) and output layer. Recurrent neural networks (RNNs) comprise memory cells. They take into account the data currently being processed, as well as the data that has come before. Currently, the most common type of RNNs is Long short-term memory (LSTM), introduced by Hochreiter and Schmidhuber [87]. LSTMs excel at predicting and classifying time series as in speech

recognition [88] and text processing [89]. A CNN with seven layers, one input layer, two convolutional layers, two max pooling layers, and two fully connected layers is shown in Fig. 2.7. Fig. 2.8 also shows a LSTM architecture as proposed by [87]. [90]

Figure 2.7: Structure of the convolutional neural network [91].

Figure 2.8: A network with eight input and four output units, and two memory cell blocks of size 2 [87].

One advantage of deep learning is that if sufficient data is provided, it can learn a hierarchical feature representation automatically. Therefore, it can be a powerful method to process the large amount of EEG data. Deep learning has shown impressive performance in several areas such as image classification [92, 93]. It also has had promising applications in biology and medicine, for instance

in plant-phenotyping [94], analysis of ECG signals [95], assessment of ultrasound images of breast lesions [96], classification of skin lesions [97], or diagnosing breast cancer [98]. [90]

Tjepkema *et al.* [90] considered 50 EEGs from focal epilepsy patients annotated by one neurophysiologist. They implemented different combinations of convolutional and recurrent neural networks, providing the probability for the presence of IEDs. They implemented deep neural networks using 5 main types: 1D CNNs, 2D CNNs, LSTMs, 1D CNNs with LSTMs and 2D CNNs with LSTMs. The network giving largest area under the ROC curve (AUC) for the test set, was validated on 7 EEG recordings with focal epileptiform discharges and 12 normal EEGs. Their final network achieved an AUC of 0.94. For validation, 47.4% sensitivity and 98.0% specificity were achieved, with FPR of 0.6/min. For the normal EEGs in the validation set, the specificity was 99.9% with FPR of 0.03/min. They concluded that deep learning can accurately detect IEDs from scalp EEG.

Xuyen *et al.* (2018)[99] applied deep belief network (DBN) for spike detection proposed by Hinton *et al.* [100]. DBN is considered as one of the most breakthrough models constructing the foundation for deep learning; and it consists of two types of neural layers: Belief Network and Restricted Boltzmann Machine, shown as in Fig. 2.9.

Deep Belief Networks could learn pattern in data even when no labeled sample is available. DBN efficiently models the generative distribution of input data. However, when used in classification tasks such as EEG classification, one needs to augment the architecture of DBN for classification accuracy. To carry out classification, a discriminative objective function on top of the existing DBN is added. There are several possible methods for classification. Firstly, one can apply standard discriminative methods which use features (outputs) generated by DBNs as inputs, for example, k-Mean, kNN, logistics regression, SVM [101]. However, a more natural way to add classification capability to DBNs is to di-

Figure 2.9: A typical DBN which contains 2 Belief Nets and 2 RBMs [99].

rectly modify the generative DBN model into a discriminative DBN model [102].

A performance comparison between using the DBN and other learning models was provided via numerical study by simulation. The compared the ANN, deep autoencoder (DAE), support vector machine (SVM) and K-nearest neighbor (kNN). In particular, the ANN is organized by an input layer, two hidden layers and an output layer followed the way of Liu [40] and Dao [103]. The DAE which is a deep generative model is modified into a discriminative model to be aiming to predict epileptic spikes that is composed of three stages including encoder, decoder and softmax layer [104]. They observed that the quality evaluation of the DBN is better than that of other models.

2.2.8 Artifact rejection

The methods described in this section employ various techniques of artifact rejection in order to distinguish IEDs from artifacts.

2.2.8.1 A knowledge-based system for artifact rejection

Glosser *et al.* [57] developed a knowledge-based system for the elimination of false positives in the automated detection of "epileptogenic sharp transients" in the EEG. Their system exploited spatial and temporal context information available on 16 channels of EEG, EKG, EMG, and EOG. A knowledge-based implementation was applied to allow the contextual rules to be expressed and refined. The resulting system was shown to be capable of rejecting a wide variety of artifacts commonly found in EEG recordings. Their hierarchical solution strategy is as follows:

- Checking appropriate channels and assess the level of consciousness (e.g., sleep stage).
- Scanning each channel for ST candidates based on their shape (e.g., electrographic parameters within proper limits).
- Checking for temporal context information around each ST candidate (e.g., presence of slow wave).
- Checking other channels for concurrent supporting evidence (e.g., presence of ST candidate in adjacent EEG channel) or contradicting evidence (e.g., presence of eye blink in EOG channel).
- Including learned knowledge from previously found ST's or artifacts
- Making a decision about the candidate ST. This is a graded decision and can be revised based on further learning as the detection proceeds.

They concluded that the existing context-based rules reject artifacts that could be false positive detections.

2.2.8.2 Independent component analysis

Independent Component Analysis (ICA) is an algorithm for signal source separation. Using statistical methods, it can separate multi-channel EEG recordings to some source signals and their scalp distribution. This method can be employed before applying traditional methods to the data, in order to separate electrocortical patterns from artefacts to some extent. In addition, giving some information on the scalp distribution of the waveforms, it helps distinguishing between different kinds of paroxysmal activity. ICA is aimed at representing the data as a linear combination of an ensemble of initially unknown and assumed to be statistically independent sources [105, 106].

ICA has been applied in combination with quadratic classifier for IED detection by De Lucia *et al.* [12]. There are many different ways or principles for quantifying the degree of independence for a set of variables, and algorithms that optimize them. Therefore, several implementations have been proposed for ICA [107–110].

In their study, they consider a fixed point iteration algorithm (FastICA), with the purpose of maximizing the negentropy of the variables; negentropy is an information-based measure of independence [108, 109]. In addition to different principles and algorithms, there are also different ways to apply ICA to physiological data [111].

Temporal ICA (tICA) is the most common approach for EEG signals. In this approach, signals at each time are decomposed into temporally independent sources having fixed spatial projections across time. The maximum number of independent components equals the number of EEG channels. tICA is a good method for signals such as eye blinks, eye movements and muscle activity [112–114], because their underlying sources are spatially compact and independent from the rest of the activity recorded by the EEG.

De Lucia *et al.* [12] designed and applied a novel variant of ICA [115], as for IED detection, tICA is problematic. The problem is that more than one

independent component are expected to account for the epileptic activity, and different patients have different foci in different locations.

De Lucia *et al.* expected that the form of IEDs are almost independent of their electrode, therefore they allowed all the IEDs to be accumulated. Then, they used a representation in which the activations of the components are sparse and independent across the time steps in the IED window. This method can characterize abrupt changes in the amplitude. Fourier-like or PCA-based representations perform lower in this respect, since they are less temporally localized [116].

The ICA decomposition was performed using Fastica [108, 109]. The independent components were fixed for both classes, and their projections onto X^k were considered as features. In the next step, they trained a quadratic classifier using these projections [117].

They then generalized the algorithm to detect a further class which included eye movements and eye blinks. They fitted a mixture of Gaussians models for different classes and computed the posterior probabilities of each class [117]. The performance of algorithm was compared using raw temporal data, as well as Principal Component Analysis (PCA). ICA gave best performance compared to raw data and PCA.

Using the ICA-algorithm for IED detection, the best performance for all splits was obtained for the case that only 3 components were extracted. The 3 components are shown in Fig. 2.10. This figure shows example of the 3 independent components extracted in the training stage after reducing data dimension to 3 using PCA. These selected components are related to the salient features that human experts consider to detect an IED. The red one is responsible for the peakedness, the green component reflects the IED duration, and the blue one represents the rebound that characterizes IEDs specially.

Figure 2.10: Example of the 3 independent components extracted in the training phase after reducing dimension to 3 using PCA [117]. (TF = time frames)

2.2.8.3 Dipole modeling

Dipole modelling is a method that helps us determine the sources underlying the changes in the electrical field of the EEG signals. This method estimates the location, orientation, and strength of each of the sources as a function of time. Dipole modelling of each individual event has been employed in [118] to help distinguish electrocortical events and artefacts, which is helpful for IED detection.

The method applied by Van Hese *et al.* [118] includes 5 steps: (1) dipole source analysis in a moving window, (2) automatic selection of focal brain activity, (3) dipole clustering to identify the epileptiform cluster, (4) deriving a spatio-temporal template of the epileptiform activity, and (5) template matching. The details of this study is as follows: The EEG segment $W_k \in \mathbf{R}^{m \times L}$ (m electrodes, and L time samples), is decomposed by PCA:

$$W_k = U_k \Sigma_k V_k^T, \quad (2.20)$$

and the first left eigenvector representing the potential topography of the first principal component, is analysed using a dipole model by minimizing the cost function:

$$PRE_k = ||\mathbf{U}_k(:, 1) - \mathbf{L}(\mathbf{r}_k)d_k\mathbf{e}_k||_2^2, \quad (2.21)$$

where $\mathbf{r}_k$ is the dipole position, and $\mathbf{d}_k = d_k\mathbf{e}_k$ is the dipole components, i.e., dipole intensity and orientation, and $\mathbf{L} \in \mathbf{R}^{m \times 3}$ denotes the lead field matrix relating the dipole parameters to the surface potentials, and PRE_k is the relative residual energy, i.e., the ratio of energy that cannot be explained by a dipolar field. For each window k , S_k is defined as the ratio of the first principal component energy:

$$S_k = \frac{\sigma_{1k}^2}{\sum_{i=1}^m \sigma_{ik}^2}. \quad (2.22)$$

Each EEG segment $\mathbf{W}_k$ is thus modelled as

$$\begin{aligned} \mathbf{W}_k &\cong \mathbf{U}_k(:, 1)\sigma_{1k}\mathbf{V}_k(:, 1)^T \\ &\cong \mathbf{L}(\mathbf{r}_k)d_k\mathbf{e}_k\sigma_{1k}\mathbf{V}_k(:, 1)^T. \end{aligned} \quad (2.23)$$

The dipole time series are contained in the first right eigenvector $\mathbf{V}_k(:, 1)^T \in \mathbb{R}^{1 \times L}$.

Focal brain activity is identified by selecting all windows k that fulfill the criteria (i) $S_k > \theta_s$, ensures the presence of a single dominant source; (ii) $PRE_k < \theta_{PRE}$, which confirms the detection of a source characterized by a dipolar field and thus a focal source. However, it is possible to obtain low RRE values for spatially extended sources [119], hence, a low RRE value is a necessary but not a sufficient condition; and (iii) $||\mathbf{r}_k||_2 < \theta_e$, sets a restriction on the dipole eccentricity such that only physiological meaningful dipole sources are kept and certain artifacts are removed [120]. Each selected window is characterized by a dipole position $\mathbf{r}_k$, orientation $\mathbf{e}_k$ and time series $\mathbf{y}_k$. The set of dipole positions and orientations is clustered using a sequential clustering algorithm. Clusters are then employed in template matching. Fig. 2.11 shows the clustering results

of the dipole position and orientation vectors that represent focal brain activity. The dipoles in frontal, top and side view are shown in the three figures at left. Each cluster is shown by a circle with a radius of 15 mm; and its center is the mean of all dipole positions of that cluster. The epileptiform cluster is plotted in red, and other clusters are plotted in blue. N denotes the total number of selected windows (and dipoles). N_i denotes the number of dipoles in cluster i . The figures at right represent the dipole time series of the epileptiform cluster with the red bold line showing the average time series [118]. This method allowed

Figure 2.11: Clustering results of the dipole position and orientation vectors representing focal brain activity [118].

automated localization of the epileptogenic focus, and good agreement with the region indicated by the experts.

2.3 Comparison of IED detection methods

In this section, we provide a comparison between the methods that have been applied for IED detection in the previous studies.

We list the methods which have been employed for IED detection together with their performance characteristics as reported by authors, in Table 2.1. In addition, we provide the size of EEG recording, and number of IEDs for training and testing in Table 2.2. In both tables, only the first author’s name has been included for brevity.

Table 2.1: Algorithms applied for IED detection, and their performance characteristics reported by authors.

Study	Method	Sensitivity	Specificity	Precision	FPR
Gotman 1976[121]	M				0.33
Gotman 1979[47]	M				0.11
Fischer 1980[32]	P, TM	0.73			1
Oliveira 1983[46]	M	0.65			
Gotman 1991[59]	SA, M				0.25
Gotman 1992[60]	M, SA			0.72	0.79
Gabor 1992[78]	NN	0.942		0.791	1.5
Hostetler 1992[48]	M	0.76		0.89	5.2
Sankar 1992[41]	P			0.175	
Dingle 1993[44]	M, ES	0.67	1		0
Pietila 1994[49]	FT, M	0.31	0.33		0
Webber 1994[76]	M, NN	0.736		0.736	6.1

Study	Method	Sensitivity	Specificity	Precision	FPR
Kalayci 1995[80]	WT, MM	0.908	0.932		
Senhadji 1995[42]	P, WT	0.86			6.8
Benlamri 1997[50]	M				
Feucht 1997[64]	HT, NN	0.86-0.89		0.846	1.83
Park 1998[82]	WT, NN, ES	0.97		0.895	
Ozdamar 1998[72]	NN	0.93		0.94	
James 1999[53]	M, NN	0.553		0.82	0.12
Sartoretto 1999[122]	WT	0.7		0.52	1.28
Ramabhadran 1999[123]	M, ES	0.96			0.4
Wilson 1999[75]	M, NN	0.47	0.9963		2.5
Black 2000[58]	M, ES				0.0068
Goelz 2000[124]	WT	0.84		0.12	
Ko & Chung 2000[73]	NN				
Kurth 2000[74]	NN	0.623		0.61	
Liu 2002[40]	P, WT, NN, ES	0.9		0.936	
Sugi 2002[56]	FT, M	0.37			16
Castellaro 2002[79]	FT, NN				
Flanagan 2003[120]	M, DA				
Latka 2003[67]	WT	0.7		0.67	
Flanagan 2003[120]	M, DA				
Adjouadi 2003[54]	M, WaT	0.82	0.93	0.92	
Adjouadi 2004[55]	M, WaT	0.79	0.88	0.85	
Acir 2005[62]	M, NN	0.891		0.859	0.125

Study	Method	Sensitivity	Specificity	Precision	FPR
Argoud 2006[81]	WT, NN, ES	0.709	0.9999		0.13
Exarchos 2006[51]	FT, M	0.861	0.917	0.831	
Tzallas 2006[77]	M, FT, NN, ES	0.892	0.911	0.825	
Inan 2007[125]	M, NN, C	0.604	0.987	0.818	
Van Hese 2008[118]	DA	0.92		0.77	
Indiradevi 2008[13]	WT	0.917	0.893	0.893	
De Lucia 2008[12]	ICA, QC	0.65	0.86		6
Nonclercq 2012[126]	TM, C	0.906		0.899	
Lodder 2013[33]	ICA, TM, SVM	0.9			2.36
Liu 2013[61]	NLEO, M, AdaBoost	0.94 ± 0.04	0.89 ± 0.07		
Zhang 2013[127]	M ^a , AdaBoost	0.956	-		
		0.961	0.98		
Chavakula 2013[128]	WT, ML	0.43-0.78	0.64-0.96		
Lodder 2014[129]	TM	0.95			
Janca 2015[130]	HT, MLE	0.97		36.3	8.2
Parthiban 2015[131]	WT, M ^a , GA	0.71		0.71	
Carey 2016[85]	NN	0.826		0.726	2.34
Tjepkema 2018[90]	Deep learning: CNN,RNN LSTM	0.47	0.98		0.6
Anh-Dao 2018[103]	WT, SF, NN	0.99			
Xuyen 2018[99]	Deep leaning	0.873	0.978		

Study	Method	Sensitivity	Specificity	Precision	FPR
Fukami 2018[132]	(DBN), WT Eigenvalue analysis, C	0.708		0.39	

M: Mimetic, P: Parametric, TM: template Matching, SA: State Analysis, NN: Neural Networks, ES: Expert System, FT: Fourier Transform, WT: Wavelet Transform, HT: Hilbert Transform, DA: Dipole Analysis, WaT: Walsh Transform, C: Clustering, ICA: Independent Component Analysis, ML: Machine Learning, Cl: Classifier, QC: Quadratic classifier, SVM: Support vector machine, GA: genetic algorithm, SF: Spectral filter, HT: Hilbert Transform, MLE: maximal likelihood algorithm, CNN: convolutional neural network, RNN: recurrent neural networks, DBN: deep belief network,

^a Sequence Merging, ^b Morphological filter

Table 2.2: Characteristics of EEG datasets considered in previous studies for IED detection.

Study	Number of subjects	Number of IEDs	Record length (min)	Number of experts	Other
Gotman 1976 [121]	93	-	2	1?	
Gotman 1979 [47]	24	-	12240	1	
Fischer 1980 [32]	10	341/341	8	7	
Oliveira 1983 [46]	5/5	-	4.2/4.2	8	
Gotman 1991 [59]	14/6	-	1400/600	2	
Gotman 1992 [60]	20	-	2000	2	
Gabor 1992 [78]	5	652	63.8	1	
Hostetler 1992 [48]	5	1393	100	6	

Study	Number of subjects	Number of IEDs	Record length (min)	Number of experts	Other
Sankar 1992 [41]	11	-	29.5	1	
Dingle 1993 [44]	11	462	180	1	
Webber 1993 [45]	10	927	40	8	
Pietila 1994 [49]	6	-	360	2	
Webber 1994 [76]	5/5	927	40	1	
Kalayci 1995 [80]	5	1200/2414	75	2	
Senhadji 1995 [42]	1?	982	10	1	
Benlamri 1997 [50]	137	-	-	1?	
Feucht 1997 [64]	4	1509	120	2	
Park 1998 [82]	32	-	-	2	
Ozdamar 1998 [72]	5	-	75	3	
James 1999 [53]	35/8	3096/190	856/192	2-3	
Sartoretto 1999 [122]	10	166	79	1	
Ramabhadran 1999 [123]	6/18	-/982	90/270	1	
Wilson 1999 [75]	50	1952	143	5	
Black 2000 [58]	521	-	10380	2-3	
Goelz 2000 [124]	11	298	278	3	
Ko 2000 [73]	20	300	-	1	
Kurth 2000 [74]	4	54/-	-	2	
Liu 2002 [40]	81	6048	~48000	1	
Sugi 2002 [56]	11	77	8	1	
Castellaro 2002 [79]	-/50	1084/-	325/-	4	
Flanagan 2003 [120]	6/20	17302/23619	6336/16920	1	

Study	Number of subjects	Number of IEDs	Record length (min)	Number of experts	Other
Latka 2003 [67]	4	340	-	1	
Pang 2003 [133]	3/10	-	-	1	
Adjouadi 2003 [54]	10/21	319	~800	3	
Adjouadi 2004 [55]	9/9	47/139	~450	2	
Acir 2005 [62]	19/10	216/93	210/228	2	
Argoud 2006 [81]	7	6721	-	3	
Exarchos 2006 [51]	25	137/137	~375	2	
Tzallas 2006 [77]	15/10	163/111	~225/150	2	
Inan 2007 [125]	5/3	53/15	-	2	
Van Hese 2008 [118]	8	1625	130	2	
Indiradevi 2008 [13]	22	684	-	2	
De Lucia 2008 [12]	7	-	121	1	
Nonclercq 2012 [126]	14	27500	575±42.5	3	Different sleep-wake cycle
Lodder 2013 [33]	8/15	482/241	175/306	-	
Liu 2013 [61]	15	142	-	-	
Zhang 2013 [127]	32+10	415+7195	1440+200	2	24 hr recording, & light sleep
Chavakula 2013 [128]	55	-	221-313	2	5 sets of different sleep-wake cycle
Lodder 2014 [129]	-	2973	175	1	
Janca 2015 [130]	30	1956	-	3	
Parthiban 2015 [131]	23	-	54960	-	

Study	Number of subjects	Number of IEDs	Record length (min)	Number of experts	Other
Carey 2016 [85]	6	-	300-1440 per patient	1	Leave-one-out CV
Tjepkema 2018 [90]	50	1478/337	-	1	
Anh-Dao 2018 [103]	12/5	-	226.1/115.6	1	
Xuyen 2018 [99]	19	1491		1	
Fukami (2018)[132]	8	-58.6±36.2 each	37.5±32.0	1	

By comparing the IED detection methods, as well as their performance characteristics listed in Table 2.1, we observed that:

1. It is difficult to compare the performance of IED detection algorithms, since different methods of statistical analysis are employed in different studies.
2. Most of the studies report only few of the common metrics for IED detection performance. In addition, many studies do not give detection rate per hour. The problem is that it is possible for a system with high values of sensitivity and specificity to have a high false detection rate per hour, when the frequency of IEDs in the EEG is high.
3. Different definitions have been considered for false and missed detections. Based on what fractions of interpreters annotate IEDs, they are defined differently in different studies. Many of them consider a waveform as an IED only if all the interpreters annotate it as IED, as in [62, 64, 74, 81, 82, 125]. Some of them annotate the paroxysmal events as IED if at least one out of a number of experts annotate it as an IED [54]. Other studies require certain fraction of interpreters to annotate events as IEDs.
4. The frequency of occurrence of IEDs in each recording, and how epileptiform an IED is, are not considered in many studies while calculating the performance metrics. Most studies consider an IED or not-IED categorization. Two studies employ a

dichotomous annotating for IEDs, definite or questionable [53, 118]. In one study a numeric scale is applied to annotate how epileptiform each event is [134]. This approach can be better, since the performance of the algorithm is judged by quantity of the detected IEDs as well as their quality.

5. It is helpful to include the frequency that IEDs happen, in the statistical analysis. Detection in records with numerous IEDs is easier than in records with very few IEDs. Hence, considering this factor can affect the results. One study suggests to include the frequency of IEDs in each record while calculating the performance, since the record with large number of IEDs would skew the overall results [135].

From Table 2.2, we observe that:

1. In some methods, individual IEDs were not annotated by the EEGers, and only whether the entire EEG contained IEDs was determined, as in [58].
2. In most studies with a large number of EEG recordings, the number of interpreters is not large, presumably to minimize study cost.
3. Many studies do not investigate different datasets for training and testing. Leave-one-out cross-validation method has been applied in some recent studies like [12, 118] in order to compensate for it.

In summary, it is not feasible to precisely compare IED detection methods due to reasons including: (i) the datasets utilized for IED detection are of various size and type; (ii) number of experts annotated them is small and different for each study; (iii) reported performance characteristics are not complete for most studies.

Moreover, there is no method with the desired performance characteristics, which is implemented on a sufficiently large dataset of separate patients for training and testing, by using enough number of experts.

Current methods are not validated on large scalp EEG datasets, and therefore are not universally accepted. The biggest obstacle for developing an efficient and reliable

IED detection algorithm is the lack of a sufficiently large database IEDs annotated by experts. So far no algorithm has been able to overcome these challenges and to offer expert-level detections, mainly due to the costs required to collect a sufficiently large and diverse database for training and validation, leading to physicians mistrusting the published algorithms.

2.4 Inter-rater agreement analysis methods

In this section, we present the studies in the literature which have investigated the agreement between experts in the interpretation of interictal EEG. Disagreement between EEGers in annotating IEDs is known to be substantial [134], and the clinical skill of EEGers varies considerably as a result of different levels of experience and education, for instance, a history of fellowship training. EEG interpretations differ considerably even among expert EEGers [7].

We present the results of the studies which analyzed expert inter-rate agreement, as follows:

Guedes de Oliveira *et al.* [46] observed poor agreement among 8 readers interpreting ten 50-s interictal EEG recordings.

In a study by Holster *et al.* [136], an IED detection algorithm is compared with a group of 6 EEGers. An expertise factor is assigned to each of them based on their experience, and they scored the IEDs in 6 records. Probability of each event annotating as IED by each scorer was calculated accordingly, and the performance of each one was compared with the algorithm. They concluded that the algorithm is not as specific as EEGers, but it is more effective than one EEGer in terms of sensitivity, due to high degree of variability between expert scoring.

In the study by Webber *et al.* [45], 8 experts interpreted 12 short EEG recordings. In their study, only 18% of all IEDs were annotated by all annotators, and 38% were annotated by only one expert [3, 45].

In another study by Wilson *et al.* [134], 5 experienced EEGers detected IEDs in 40 epilepsy patients and 10 control subjects. Spike perception modelling was based on dichotomous values as well as continuous values. They concluded that the reliability

is higher using continuous values. The average inter-reader correlation is reported to be 0.79.

In a study by Black *et al.* [58], a large set of unselected recordings was analyzed by multiple readers. Inter-rater agreement was analyzed for 521 twenty-minute EEGs. For 106 cases assessed by 3 experienced EEGers, the 3 agreed on the presence of IEDs in only 39% of the EEGs for which at least one reader annotated IEDs. For 415 cases in which two readers analyzed the same EEGs, agreement was 55%. This substantial lack of agreement in an unselected recording indicates the risk of selection bias when EEGs are chosen by experts for an IED detection study.

Halford *et al.* [7] also reported that inter-rater agreement was only moderate (Fleiss kappa = 0.43) among 11 clinical neurophysiologists marking a set of 2571 selected IEDs of 100 EEG recordings.

In the study by scheuer *et al.* [137], the human readers showed only a fair level of agreement, with an average of approximately 45% sensitivity for each another's sIED annotations.

In summary, previous studies suggest that: (1) there is a low degree of agreement between experts in IED interpretation; (2) using computer algorithms can increase sensitivity and uniformity in IED detection, compared to using one expert; (3) it is more reliable to annotate IEDs using scales having more than only two IED and non-IED categories.

In this chapter, we discussed and explored several techniques applied for IED detection. We also discussed the studies on expert agreement analysis. In the next chapters, we elaborate the techniques performed in this thesis towards automated detection of IEDs.

Chapter 3

A Fast Automated Method to Facilitate the Detection of Interictal Epileptiform Discharges in the Scalp Electroencephalogram

In this chapter, we propose and explain an algorithm for rejection of background data from EEG. This algorithm is designed to eliminate background waveforms from EEG data, and consequently speed up the IED detection process. We also describe the dataset employed in this study, as well as the features, the classification methods applied for IED detection combined with the background rejection algorithm, and validation of this technique.

3.1 Introduction

As explained in Chapter 2, there is currently no IED-detection software program available which is commonly-accepted to aid in diagnosis of epilepsy in clinical applications. We also mentioned in Chapter 1 that one common problem with many methods for automated IED detection is the lack of a sufficiently large database containing various IED profiles

for training and validating the performance of IED detection algorithms. Hence, they cannot be used reliably for clinical applications.

In this study, we consider a dataset of routine scalp EEG recordings of 93 epileptic and 63 healthy subjects. In order to deal with an immense volume of EEG data, we develop an ensemble system comprising simple classifiers for rejecting most background data. The waveforms remaining from this algorithm include almost all IEDs, together with a low proportion of background waveforms which are similar to IEDs. A similar approach was taken for face detection by Viola *et al.* [16], in which they filter out over 50% of the images while preserving 99% of the faces, as a part of their face detection algorithm.

In the remainder of this chapter, we describe the process of designing the ensemble system comprising simple classification units, the feature extraction, feature selection, training and evaluating the cascade. Finally, we demonstrate how employing this algorithm can speed up IED detection.

3.2 Materials and Methods

In this section, we describe the overall idea of the background rejection algorithm and explain the EEG dataset considered in this study. The purpose of this study is to develop a fast algorithm to reliably remove non-IED (“background”) waveforms from EEG. This is particularly useful with large EEG datasets. This algorithm can serve as the initial step of an IED detection system. This algorithm is a feature-based ensemble method for IED detection which detects IEDs by excluding the backgrounds. For this purpose, we need to select the best EEG features, which give highest discrimination between IED and non-IED waveforms, and we employ these features to reject background waveforms in consecutive steps. Therefore, we create a large library of features in order to pick the best ones.

3.2.1 Scalp EEG data

We consider in this study EEG recordings consisting of clinically-obtained scalp EEG recordings at the Massachusetts General Hospital (MGH) from 156 subjects with a total

length of 4454.2 minutes, with an average of 28.55 minutes per EEG recording. Retrospective analysis of the data was performed under an IRB-approved protocol. Among the 156 subjects, 93 are epileptic patients whose EEGs contain IEDs, and 63 are healthy subjects with IED-free EEGs. In most IED analysis studies, only EEGs from patients with epilepsy are analyzed. By contrast, here we consider the EEG from subjects without IEDs for both training and validation. The data was acquired by means of the international 10-20 system of electrode placement, and it is resampled to 128Hz. A notch filter is applied at 60 Hz for removing power line interference. Baseline drift is removed by applying a high-pass filter by cut-off frequency of 0.1 Hz. Data is cross-annotated by two clinical neurophysiologists from MGH using NeuroBrowser software [138]. In total, there are 18,164 IEDs in the dataset. We apply 5-fold cross-validation to split the data into training and testing sets, and subjects are randomly assigned to each fold. The data from each subject is only used either for training or testing for each fold. In each round of cross-validation, we train the system on 4 folds, and test it on the 5th fold.

3.2.2 Data preparation and feature extraction

We propose to utilize various quick-to-compute features in consecutive steps to eliminate background waveforms from the interictal EEG. Therefore, we set out to select features which yield highest discrimination between IED and non-IED waveforms. For this purpose, we compare a library of features based on the proportion of background waveforms each of them can reject. In the remainder of this section, we explain how the features are extracted, selected, and how the system is trained, tested, and evaluated.

3.2.2.1 Data preparation

In this study, we define IEDs as waveforms that are annotated by both annotators as IED. We define backgrounds as waveforms that do not overlap with annotated IEDs. We do not consider as background the waveforms that overlap in time with any annotated IED in any channel. We consider 0.5s windows for IEDs and background waveforms for this analysis.

3.2.2.2 Feature extraction

We compute several feature categories for each 0.5s EEG segment. We consider a large set of EEG features, allowing us to select a suitable discriminative EEG feature at each step in the cascade. The categories of features and their subgroups and parameters are listed in Table 3.1. The choice of features is based on the following criteria: (1) computational complexity, (2) availability of methods to compute the features, (3) literature suggesting certain features are useful for IED detection or similar applications, (4) neurologists' opinion about usefulness of certain features.

We performed all data processing and statistical analysis offline in Matlab (Matlab 2014a, The Math Works Inc., Natick, MA, 2014). We now explain the features.

Table 3.1: Summary of all computed feature categories and their subgroups and parameters.

Feature	Parameters/subgroups
Waveform morphological features	Peak voltage, Minimum voltage, Peak-to-peak voltage, Slope, Duration, Line length
Nonlinear energy operator	Resolution parameter k=1 to 40
Wavelet analysis	Continuous wavelet transform coefficients 53 wavelets, Scaling parameter s =1 to 20
	Discrete wavelet transform coefficients 53 wavelets, 4 decomposition levels
EEG main frequency bands	Delta, Theta, Alpha, Beta, Gamma
Spectrogram features	Mean PSD values, Maximum PSD, Frequency and Time of maximum PSD, Standard deviation of PSD values

3.2.2.3 Waveform's morphological features

Morphological features computed for each 0.5 s segment include: (a) peak voltage, (b) minimum voltage, (c) peak-to-peak voltage, (d) slope, (e) duration, and (f) line length. We calculate the peak voltage as the maximum voltage of the EEG segment, the peak-

to-peak voltage as the difference between maximum and minimum voltage values, the duration as the time difference between maximum and minimum voltages, and slope as peak-to-peak voltage divided by duration.

We compute the line length of an N -point signal $x(n)$ as [139]:

$$L = \sum_{i=1}^N |(x(k-1) - x(k))|. \quad (3.1)$$

3.2.2.4 Nonlinear energy operator

NLEO is a useful technique for detection of the sudden occurrence of high frequency signals, and it is commonly utilized in signal processing, image processing and AM/FM demodulation [140, 141]. NLEO accentuates high-frequency content, and due to the instantaneous nature of NLEO and having low computational burden, it is an appropriate feature to be applied in IED detection [142]. Some previous studies have applied the nonlinear energy operator (NLEO) for IED detection [61, 142]. NLEO is computed as:

$$\phi_k\{x[n]\} = x^2[n] - x[n-1]x[n+1], \quad (3.2)$$

where $x[n]$ denotes the signal at time n . NLEO is sensitive to high noisy peaks, especially in situations with low signal to noise ratio (SNR). Choi and Kim [143] proposed the extended version k-NLEO defined as:

$$\phi_k\{x[n]\} = x^2[n] - x[n-k]x[n+k], \quad (3.3)$$

where k is the resolution parameter which is related to the peak width of the detected IED candidates and makes NLEO sensitive to the frequency. It also makes it relatively immune to out-of-band noise [143]. We chose k to have values in the range of 1 to 40. We first calculated the NLEO value for every sample point of the EEG signal, and next we extracted segments from the EEG signals. We considered the standard deviation and maximum of the NLEO values as the features for each segment.

3.2.2.5 Wavelet analysis

The wavelet transform is beneficial for characterizing several aspects of EEG signals such as non-stationary transient events [3]. We provided detailed explanations and definitions for wavelet transform in Chapter 2 (Section 2.2.5).

For this study, we applied multiple wavelet transforms in order to generate features. We tested several wavelet basis functions, since it is unclear a priori which wavelets are best suited to detect IEDs. We calculated 53 wavelets including symlets (sym), coiflets (coif), Daubechies (db), biorthogonal (bior), reverse biorthogonal (rbio), and discrete approximation of Meyer (dmey) wavelet basis families with different degrees [69].

In the continuous wavelet transform (CWT), inner products are applied for measuring the similarity between a signal and the wavelet ψ . Specifically, the signal is compared to shifted and compressed or stretched versions of ψ :

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right), \quad (3.4)$$

where u and s are translation and scaling parameters, respectively. By varying the values of these parameters, the CWT coefficients $C(u, s)$ are obtained. In addition, depending on what signal features we are trying to detect, we can select a wavelet that facilitates the detection of that feature.

The continuous wavelet transform $C(u, s)$ is derived from the mother wavelet $\psi_{u,s}(t)$ (Equation (3.4)) by shifting and scaling the wavelet basis function as:

$$C(u, s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} f(t) \psi_{u,s}(t) dt. \quad (3.5)$$

The scale factor determines compression or stretching. We tested different values for the scale factor s , ranging from 1 to 20. We computed features as the maximum absolute value and standard deviation of the coefficients calculated on each segment, for the wavelet transform computed by each of the wavelet basis functions and various scaling factors.

We also calculated the discrete wavelet transform (DWT). In DWT the signal is decomposed into multiple scales. In Fig. 3.1, we illustrate the wavelet decomposition

of signal $s[n]$, where Lo and Hi are high-pass and low-pass filters, respectively. The parameter D_i is the detail, whereas A_i represents the approximation at the i th level. These coefficients are obtained by convolving the signal by low-pass and high-pass filters, followed by a down-sampling step. Approximation coefficients of previous level are utilized to calculate the coefficients of current level. We illustrate an IED together with its wavelet coefficients at various levels for a Daubechies 3 (DB3) wavelet in Fig. 3.2.

Figure 3.1: Subband decomposition for implementation of discrete wavelet transform.

Figure 3.2: Wavelet decomposition of a sample IED.

Four levels of decomposition were constructed for the discrete wavelet transform; the detail and approximation coefficients were derived for every level [69]. These coefficients were then reconstructed for every level, yielding an output of the same length as the input vector. We first calculated the wavelet coefficients for each EEG recording as a whole, and then we extracted the coefficients for every 0.5s segment. This helps to avoid distortions at the boundaries of the segments and reduces the computational complexity. We calculated the features based on the resulting DWT coefficients, which include standard deviation, and maximum absolute value of the detail and approximation coefficients at every level. We calculated them within each segment for the DWT applied using each of the wavelet basis functions. At a sampling rate of 128 Hz, the filters applied in DWT for each of the decomposition scales include the frequency ranges of d1 (32 – 64 Hz), d2 (16 – 32 Hz), d3 (8 – 16 Hz), d4 (4 – 8 Hz), a1 (0.5 – 32 Hz), a2 (0.5 – 16 Hz), a3 (0.5 – 8 Hz), and a4 (0.5 – 4 Hz).

3.2.2.6 Various frequency bands

In addition to extracting the above mentioned features for signals filtered between 0.5 to 128 Hz, we computed them on bandpass filtered signals in the standard EEG frequency bands defined as Delta (<4 Hz), Theta (4 – 8 Hz), Alpha (8 – 12 Hz), Beta (12 – 32 Hz), and Gamma (>32 Hz) [4]. We considered the combinations of these bands as well.

3.2.2.7 Spectrogram and power spectral density

We computed the spectrogram of the signal by applying the short time Fourier transform (STFT) [144] with Hamming windowing. We defined several features based on the PSD values for each time segment of the signal: (a) mean of the PSD values, (b) maximum PSD value, (c) frequency and time of the maximum PSD value, (d) standard deviation of PSD values.

In summary, we computed features which represent signal morphology, as well as various features to represent the frequency aspects of the signal, in the feature library. The aim is to find the most discriminant features among a large pool of automatically computed features, most of which might not be significantly discriminant.

3.2.3 Implementing the background rejection algorithm

In this section, we describe the training and testing phases of the proposed algorithm. In addition, we explain the method we applied to evaluate the algorithm in terms of reduction in the computation time, as well as IED detection methods implemented for this purpose.

Firstly, we train a cascade of simple thresholding steps by which we can retain most of the IEDs in the training set. Next, we apply the trained cascade on the test set. We show flowcharts of training and testing phases in Fig. 3.3a and Fig. 3.3b, respectively. At last, we evaluate the efficiency of the proposed algorithm by detecting IEDs by means of a number of common classification methods, with and without applying the background rejection algorithm. In the following, we explain the training and evaluation process in more detail.

Figure 3.3: Flowcharts showing (a) training and (b) testing processes.

3.2.3.1 Training

We randomly sampled background waveforms of length 0.5s without replacement from EEG of the subjects in the training set. In this manner, we extracted 5 times more background waveforms than IEDs. For EEGs with zero or a small number of IEDs, we extracted a minimum of 2000 waveforms.

Next, we calculated the empirical cumulative density function (CDF) for each feature, for the background waveforms as well as IEDs. Then we selected a threshold θ for each feature; we selected θ such that almost all IEDs were retained. We set the fraction

α of IEDs which are allowed to be missed in each step to $\alpha = 0.001$ and computed θ from the CDF of the IEDs accordingly. We then derived the ratio β of rejected background data for each feature, from the backgrounds CDF. We sorted the features based on their rejection ratio β , and we selected the feature with highest value of β for the first step. After selecting the top feature, we applied its corresponding threshold θ on the whole training set.

In this procedure, all waveforms having a value lower than the selected threshold θ for a given feature would be rejected from the dataset. Waveforms having feature values higher than the threshold will be passed to the next step.

We illustrate in Fig. 3.4 an example of the empirical CDF for the IEDs and background waveforms. The threshold θ which is selected such that $1 - \alpha\%$ of IEDs are retained, is shown as a vertical line. By applying this threshold, we eliminate $\beta\%$ of background data having feature values below θ . We apply a similar process in every step of the training phase.

We performed the same process for the following steps: we calculated all the features for the remaining segments from the previous step; then we determined the feature leading to the highest ratio β_T , and so forth. The total background rejection rate (specificity) after the L th step is:

We repeat the same process for the next steps in the background rejection cascade; for the remaining waveforms, we determine the feature with the highest ratio β , and remove more data according to its threshold, etc. The total background rejection rate (specificity) after the L th stage is:

$$\beta_T = \prod_{\ell=1}^L \beta_{\ell}, \quad (3.6)$$

and the overall sensitivity S_L is:

$$S_L = (1 - \alpha)^L. \quad (3.7)$$

The features are added until the rejection rate ceases to increase significantly or until the

Figure 3.4: The features selection process applied in each step by computing the empirical cumulative density function.

fraction of missed IEDs falls below a specific value. Thresholding on each feature can be viewed as training a weak classifier. A weak classifier performs poorly and classifies with an accuracy slightly higher than a random classifier. Therefore, each weak classifier $h(w, f, \theta)$ is defined as:

$$h(w, f, \theta) = \begin{cases} 1, & \text{if } f(w) \geq \theta \\ 0, & \text{otherwise,} \end{cases} \quad (3.8)$$

where f represents a feature among all the considered features, w is a 0.5s window of EEG signals, and θ is the threshold set on the feature. Only waveforms for which $h(w, f, \theta) = 1$ are retained for the next step. We illustrate the overall process for training in Fig. 3.5a.

3.2.3.2 Testing

After building the cascade, we assessed its performance on the test set. Testing waveforms were extracted from all the EEG channels, by applying a sliding window with 50% (0.25 s) overlap. By contrast, in the training phase we randomly sampled the background waveforms. For each step, we computed the feature selected in the training phase for all the waveforms and then applied the thresholds specified in the training phase. The decision for each step is defines as:

$$h(w, f_s, \theta_s) = \begin{cases} 1, & \text{if } f(w) \geq \theta \\ 0, & \text{otherwise.} \end{cases} \quad (3.9)$$

W_L is the set of test waveforms after L th stage: if $h(w, f_s, \theta_s) = 1$, then $w \in W_L$. W_{Test} will be updated by keeping only the waveforms which are classified as IED, therefore, $W_{Test}=W_L$. In other words, we updated the set of testing waveforms after every step by keeping only the waveforms which were classified as IED, i.e., the segments for which $h(w, f_s, \theta_s) = 1$.

We computed the cascade's performance as the overall specificity and sensitivity in terms of IED detection. After building the cascade, its performance was tested on the test set. Testing waveforms were extracted from all the EEG channels, by applying a sliding window with 50% (0.25 s) overlap. This is in contrast to training, for which background was randomly sampled. The gold standard considered for testing is the IED annotations by the two neurologists. For each step, the feature selected in the training phase was computed for all waveforms and then the thresholds specified in the training phase were applied. The decision for each step is defines as:

$$h(w, f_s, \theta_s) = \begin{cases} 1, & \text{if } f(w) \geq \theta \\ 0, & \text{otherwise} \end{cases} \quad (3.10)$$

W_L is the set of test waveforms after L th stage: if $h(w, f_s, \theta_s) = 1$, then $w \in W_L$. W_{Test} will be updated by keeping only the waveforms which are classified as IED, therefore, $W_{Test}=W_L$.

The cascade's performance was computed as the overall specificity and sensitivity in terms of IED detection.

3.2.3.3 Evaluating the efficiency of the algorithm

In order to assess the proposed background rejection algorithm, we evaluated two classification procedures: (i) we trained an IED detector on the training set, and applied it on the testing set, (ii) we trained the detector on the remaining waveforms of the training set which were not rejected by the background rejection algorithm; we then applied it to classify the remaining waveforms from the testing set. We illustrate these classification procedures in Fig. 3.5. In both cases, we calculated the time taken to classify the test data and the performance characteristics. Comparing the computation time and the IED detection accuracy of both methods, we can determine whether the proposed background rejection cascade speeds up the classification. We apply as classifiers support vector machine, random forest, and K-nearest neighbors. In the following, we will briefly discuss each of those classifiers.

Figure 3.5: Two procedures for IED detection (a) using a classifier and (b) applying the background rejection algorithm followed by a classifier.

- **Support Vector Machines (SVM):** Support vector machines are considered as error-based learning models. Training an SVM involves searching for the decision boundary, or separating hyperplane which leads to the maximum margin so that it best separates

the levels of the target value. An SVM can easily be trained and is not too susceptible to overfitting. In addition, an SVM is highly effective for high-dimensional data [145]. In this study, we applied SVM with Gaussian radial basis kernel.

- **Random Forest (RF):** Decision trees are very sensitive to changes in the dataset, and even a small change can result in different feature being selected to split the dataset at the root, the tree, and consequently in the subtrees. In bagging (bootstrap aggregating), which is an ensemble method, each model in the ensemble is trained on a random subset of the data, generated by sampling with replacement. These random samples are known as bootstrap samples, and one model is produced from each bootstrap sample. However, when applying bagging on decision trees, each bootstrap sample uses only a randomly selected subset of the features. This sampling of the feature set is known as subspace sampling. It encourages the diversity of the trees within the ensemble and also reduces the training time for each tree. The combination of bagging, subspace sampling, and decision trees is known as a random forest model [145].
- **K Nearest Neighbors (KNN):** The K-nearest neighbor algorithm is a similarity-based learning method in which the Euclidean distance is measured between a signal and the training examples. In 1-nearest neighbor algorithm, there is a set of local models defined from a single instance. Considering only a single instance makes the algorithm sensitive to noise. Therefore, in order to dilute the dependency of the algorithm on individual instances, the set of K nearest neighbors to the signal are considered for classification [145]. The value of K is optimized by cross-validation

We applied these three algorithms as IED detectors and evaluated the performance of the background rejection algorithm based on the detected IEDs.

3.3 Results

In this section, we explain the results of the training and testing phases of the proposed algorithm, and the combination of it with conventional IED detection methods. We present all major performance metrics including ROC curves, sensitivity, specificity, precision, and false positive rate.

We applied the feature selection process illustrated in Fig. 3.4 for multiple steps in the training phase. We show the feature value computed for a sample EEG recording, and the threshold θ applied on it in Fig. 3.6. θ is shown as a horizontal line. We observe that by eliminating the EEG segments lower than the horizontal line, a significant portion of background waveforms are rejected.

Figure 3.6: The feature $\text{DWT-rbio3.3-max}(|a_2|)$, computed for a sample EEG recording. The horizontal line shows the threshold to be applied on it.

We trained the cascade and applied it on the testing data. We list the sensitivity and specificity of the cascade after each individual step for both training and testing, in Table 3.2 and Table 3.3, respectively. We also present the features selected and applied for the 10-step cascade, in Table 3.4. We observe that the 10-step cascade rejects 78% of the background waveforms in the testing phase, while retaining 97% of the IEDs. Since this algorithm is developed for rejecting background waveforms and retaining almost all the IEDs, we have to consider a sensitivity versus specificity trade-off. Therefore, we kept the sensitivity high but allowed the specificity to have lower values.

In addition, the feature rankings obtained while training at each step, are interesting in their own right for feature selection purposes. Among the considered features,

Table 3.2: Training and testing sensitivity for every 5 fold (F1 to F5), as well as the average over all cross-validation folds.

Steps	Training						Testing					
	F1	F2	F3	F4	F5	Mean	F1	F2	F3	F4	F5	Mean
1	0.9982	0.9981	0.9983	0.9981	0.9982	0.9982	0.9986	0.9773	1	0.9992	0.9856	0.9921
2	0.9963	0.9971	0.9965	0.9962	0.9965	0.9965	0.9957	0.966	1	0.9979	0.9831	0.9886
3	0.9945	0.9891	0.9945	0.9942	0.9952	0.9935	0.9936	0.966	0.9996	0.9979	0.9828	0.988
4	0.9933	0.9873	0.9851	0.9923	0.9933	0.9903	0.9921	0.963	0.9996	0.9979	0.9817	0.9869
5	0.9857	0.9865	0.9837	0.9837	0.9916	0.9862	0.9921	0.9621	0.9996	0.9979	0.9659	0.9835
6	0.9838	0.9851	0.9818	0.9820	0.9849	0.9835	0.9896	0.9619	0.9996	0.9975	0.9659	0.9829
7	0.9818	0.9832	0.9800	0.9802	0.9830	0.9817	0.9878	0.9611	0.9996	0.9975	0.9641	0.982
8	0.9801	0.9814	0.9784	0.9783	0.9816	0.9800	0.9796	0.9606	0.9994	0.9966	0.9634	0.9799
9	0.9785	0.9795	0.9767	0.9764	0.9800	0.9782	0.9778	0.9604	0.9994	0.9966	0.9194	0.9707
10	0.9766	0.9776	0.9749	0.9746	0.9781	0.9764	0.9739	0.9599	0.9991	0.9962	0.9176	0.9693

Table 3.3: Training and testing specificity for every 5 fold (F1 to F5), as well as the average over all cross-validation folds.

Steps	Training						Testing					
	F1	F2	F3	F4	F5	Mean	F1	F2	F3	F4	F5	Mean
1	0.4924	0.6369	0.5239	0.4801	0.5262	0.5319	0.5404	0.6754	0.3797	0.6068	0.6504	0.5705
2	0.5732	0.7166	0.6203	0.5957	0.5938	0.6199	0.6326	0.7594	0.4538	0.7025	0.6975	0.6492
3	0.6201	0.7441	0.6648	0.6360	0.6388	0.6608	0.6570	0.7594	0.5159	0.7432	0.7253	0.6802
4	0.6526	0.7669	0.6948	0.6652	0.6743	0.6907	0.7109	0.7804	0.5167	0.7621	0.7458	0.7032
5	0.6789	0.7830	0.7177	0.6939	0.7025	0.7152	0.7109	0.7967	0.5447	0.7621	0.7718	0.7172
6	0.7066	0.7973	0.7314	0.7124	0.7272	0.7350	0.7405	0.8071	0.5711	0.7761	0.7718	0.7333
7	0.7245	0.8077	0.7424	0.7265	0.7477	0.7497	0.7480	0.8183	0.5842	0.7888	0.7899	0.7458
8	0.7417	0.8161	0.7503	0.7420	0.7626	0.7625	0.7663	0.8278	0.5952	0.8043	0.8020	0.7591
9	0.7569	0.8238	0.7578	0.7551	0.7855	0.7758	0.7731	0.8364	0.6060	0.8137	0.8414	0.7741
10	0.7672	0.8304	0.7673	0.7658	0.8018	0.7865	0.7790	0.8433	0.6220	0.8249	0.8511	0.7841

discrete wavelet coefficients are the most discriminative based on the CDF, followed by morphologic features, continuous wavelet transform, nonlinear energy operator, and spectrogram features.

Moreover, the top features were among the ones computed on the high-pass filtered signals. And according to the CDFs and background rejection rates calculated for

Table 3.4: The features selected at every step of the cascade in the training phase.

Step	F1	F2	F3	F4	F5
1	DWT, rbio2.4, $\max(a3)$	DWT, rbio3.3, $\max(a2)$	DWT, rbio3.1, $\max(a3)$	DWT, rbio4.4, $\max(a3)$	DWT, db8, $\max(d4)$
2	DWT, coif3, $\max(d3)$	DWT, rbio3.1, $\max(d1)$	DWT, db9, $\max(d3)$	DWT, dmey, $\max(d3)$	DWT, dmey, $\max(a3)$
3	DWT, rbio3.3, $\max(a3)$	Peak voltage(V_p)	DWT, db10, $\text{std}(d4)$	DWT, coif4, $\text{std}(d4)$	DWT, db9, $\max(d3)$
4	DWT, coif5, $\max(d3)$	DWT, bior2.4, $\text{std}(a2)$	Peak voltage(V_p)	DWT, rbio6.8, $\max(a3)$	DWT, dmey, $\max(d3)$
5	Peak-to-peak voltage(V_{pp})	DWT, rbio3.3, $\max(d2)$	DWT, rbio1.5, $\max(a2)$	Peak-to-peak voltage(V_{pp})	DWT, coif5, $\text{std}(d4)$
6	DWT, sym7, $\max(d4)$	DWT, dmey, $\max(d3)$	DWT, sym7, $\text{std}(d4)$	DWT, rbio3.1, $\max(a3)$	Peak-to-peak voltage(V_{pp})
7	DWT, db6, $\text{std}(a3)$	DWT, rbio1.5, $\max(a2)$	DWT, sym4, $\text{std}(a2)$	DWT, rbio2.4, $\max(a2)$	DWT, db10, $\text{std}(a3)$
8	DWT, rbio2.2, $\max(d3)$	DWT, rbio5.5, $\max(a1)$	DWT, rbio2.8, $\text{std}(a3)$	DWT, coif4, $\max(d3)$	DWT, rbio3.3, $\max(d2)$
9	DWT, coif5, $\max(a3)$	DWT, bior1.5, $\max(a2)$	DWT, rbio1.3, $\text{std}(a2)$	DWT, coif5, $\max(a3)$	DWT, dmey, $\text{std}(d4)$
10	DWT, db10, $\max(a3)$	DWT, sym4, $\text{std}(a2)$	DWT, dmey, $\max(a3)$	DWT, db10, $\max(d4)$	DWT, db9, $\text{std}(a3)$

the first step, the features computed on band-pass filtered signals between 0.5-8 Hz (Delta and Theta bands), 8-12 Hz (Alpha band), 12-32 Hz (Beta band) and > 32 Hz (Gamma band) come next in the ranking.

From Table 3.4 we also observe that discrete wavelet transform coefficients are the main features selected among all other features for all folds, and only few morphological features can be observed in the list. This suggests that we can train the systems only using discrete wavelet analysis and morphological features. This can save the training time especially once we have even larger datasets.

We show the overall background rejection rates β_T after every step, for both training and testing sets, in Fig. 3.7. It was expected that the rates increase after every step. However, we observe that this increase becomes smaller after each consecutive step. In Fig. 3.8, we show the sensitivity and specificity of the cascade applied on the testing set, versus the number of IEDs in each EEG. Our purpose is to investigate how the performance of the algorithm depends on the number of IEDs in the EEG records. We observe that the algorithm has a low sensitivity for 5 patients. The algorithm's sensitivity equals zero, for four patients with 1, 1, 2, and 4 number of annotated IEDs, and it is 0.031 for one patient with 226 annotated IEDs. We depict all IEDs for the former, as well as four IEDs for the latter in Fig. 3.9, and indicate them as P1, P2, P3, P4, and P5, respectively. The index of displayed IED for each patient is shown in parentheses.

Figure 3.7: Overall percentage of rejected background after every step, for training and testing sets.

On close inspection, the probable reasons that the cascade discards some IEDs illustrated in Fig. 3.9 and cannot discriminate them from background are: (1) small amplitudes of these waveforms; (2) they do not stand out very much from the background, i.e., several of them look very similar to the surrounding background waves; (3) some are probably mislabeled, i.e., would not be universally agreed to represent IEDs (such as P2, P3(1), P4(4), P5(1,2,3,4)).

The above-mentioned factors are also considered by experts when they judge whether a waveform is “epileptiform”. In summary, it seems understandable and reassuring that these waveforms are the kinds of examples being missed by the algorithm.

Fig. 3.8 also shows that specificity value varies among the subjects. For the EEGs with low specificity, the characteristics of background waveforms seem to be similar to IEDs, and therefore this algorithm fails to eliminate them. These waveforms must be classified in the following IED detection step.

Finally, we applied the IED detection methods with and without eliminating background waveforms. In Table 3.5, we present the overall computation time on the

Figure 3.8: Sensitivity and specificity versus the number of IEDs in each EEG record for test set.

testing set for one minute of 19-channel EEG data. We measured the time for (a) background rejection cascade (as in Fig. 3.3b), (b) classifying the EEG with only SVM, KNN, or RF detector (as in Fig. 3.5a), (c) classifying by applying background rejection followed by an SVM, KNN, or RF (as in Fig. 3.5b).

Table 3.5: Overall computation time (in seconds) on the test set for one minute of 19-channel EEG data.

Method	KNN	SVM	RF
10-step background rejection		0.16	
Classifier	11.33	0.81	0.61
Classifier + 10-step background rejection	2.41	0.29	0.34
Speed-up ratio	4.69	2.74	1.76

Figure 3.9: Examples of IEDs that were missed by the background rejection cascade from 5 EEGs (P1 to P5) for which the algorithm has a low sensitivity.

We performed the analysis on a 3.4 GHz Intel Core i7 processor. In the reported time measurements, we considered the entire process including re-montaging, filtering, segmenting the data, feature computations, and applying the thresholds for background rejection, as well as applying the classifiers.

The computational time for a 10-step cascade of background rejection was 0.16 seconds. Speed-up ratio in Table 3.5 refers to the ratio by which each detection method's speed is increased. We observe that applying background rejection speeds up the detection for all considered classifiers.

We depict the ROC and precision-recall curves for the two approaches in Fig. 3.10a and Fig. 3.10b, respectively. Considering a recall of 40% or more, the difference in the precision of the two procedures is insignificant. These plots show that both ROC and precision-recall do not significantly change when background rejection is applied. Therefore, although a large amount of EEG data is removed by background rejection, we do not compromise on the performance of IED detection.

Fig. 3.11 shows the total number of detections per minute for each EEG record-

Figure 3.10: ROC and precision-recall curves for three classification methods with and without background rejection (BR).

ing. The boxplots for healthy and epileptic subjects are presented as well. We observe that the detection rate is lower for IED-free subjects. In order to verify this observation, we applied a two-sample t-test on the two groups.

Figure 3.11: Detection rate per minute versus the number of IEDs in each EEG record for test set (left), and boxplots of detection rates for EEGs with and without IEDs.

The t-test is a way to examine the difference between two groups, defined by a categorical independent variable, on a continuous independent variable. In its most basic,

one-sample form, the t-test simply asks whether the mean of a dependent variable for a single group is different from some value (usually zero). With slight modification, the two-sample version of the t-test asks whether the difference between the means of two groups is different from zero (i.e., whether the group means are equal). The two-sample version can be further modified to look at differences between two groups that are related to each other. In its essence, a t-test is a signal-to-noise ratio. The t value is a single number that captures the ratio of the “signal,” typically a mean or difference between means, relative to the “noise,” or some measure of the standard error. Hence, the most general equation for testing whether the “signal,” M , is greater than zero, is

$$t = \frac{\text{signal}}{\text{noise}} = \frac{M}{SE} \quad (3.11)$$

where M can be a mean, a difference between means, or even a difference between a mean and some hypothesized value, and SE is the corresponding standard error [146]. Statistical significance in this context is usually defined as $P\text{-value} < 0.05$. $P\text{-value}$ is the probability that a statistical summary of the data, would be equal to or more extreme than its observed value [147]. The smaller $p\text{-value}$ indicates higher significance. Consequently, the null hypothesis H is rejected if $p\text{-value}$ is less than or equal to a threshold value α that is known as the level of significance. α is not derived from any observed data and does not depend on the hypothesis; rather, it is set arbitrarily before analyzing the data. By convention, it is commonly set to 0.05, 0.01, 0.005, or 0.001.

In this study, the null hypothesis (H) is that the detection rates from healthy and epileptic subjects come from independent random samples of normal distributions having equal means and equal or unequal unknown variances. Hence, the alternative hypothesis is that the data comes from populations with unequal means. After applying this test, the null hypothesis was rejected, with $p\text{-value} = 9.98e^{-10}$. Therefore, there is a significant difference in the IED detection rates for epileptic and healthy EEGs, and the detection rate is higher for epileptic EEGs. This shows that our proposed method can successfully discriminate between epileptic and healthy EEGs which is the goal of IED detection studies.

Fig. 3.12 shows the number of false alarms per minute for each EEG recording.

The boxplots for IED-free and epileptic subjects are presented as well. We observe that the false positive rate is lower for IED-free subjects. In order to verify this observation, we applied a two-sample t-test on the two groups.

Figure 3.12: False alarm rate (false positives per minute) versus the number of IEDs in each EEG record for test set, along with the box plot for IED-free and epileptic subjects.

3.4 Discussion

In this section, we summarize the key results and compare the present study with previous studies.

In order to assess the proposed background rejection method, we combined it with generic IED detectors, and compared its performance to those IED detectors without background rejection. We observed from Fig. 3.10 and Table 3.5 that combining background rejection algorithm with other detection methods helps to significantly reduce the computation time, while maintaining the IED detection accuracy.

The results also show that SVM performs better in IED detection compared to KNN and RF, while the performance of these algorithms in terms of false positive rate and sensitivity is not yet satisfactory. However, the goal of this study is to develop a procedure for background rejection, allowing us to reduce the computation time by removing irrelevant EEG segments.

We also noticed that there are few EEG recordings for which the algorithm fails to perform as expected. We analyzed these recordings in more detail and suggested possible reasons which can be beneficial for future studies.

Additionally, rejecting more proportion of background in normal subjects compared to epileptic subjects shows that this method as a preprocessing algorithm is suitable for filtering out normal EEGs and therefore can lead to less number of false positive detections in a healthy EEG. This can be useful for subject-level EEG classification following IED detection.

Moreover, we explained how we can view the proposed method as an approach to select features for IED detection. We observed that discrete wavelet coefficients are the leading features to discriminate between EEG background waveforms and IEDs.

Most methods applied in the literature analyzed relatively small EEG datasets, or EEG datasets with a small number of subjects. For instance, the study by Flanagan *et al.* [120] explored 23,256 minutes of EEG recording; their dataset consists of only 26 subjects, though. In this study, we considered a large dataset of 156 patients with a total of 4,454.2 minutes of 19-channel EEG. Furthermore, our dataset also contains 63 IED-free EEG recordings.

In addition, we applied 5-fold cross validation in this work, in contrast to most previous studies. Many studies failed to exploit different datasets for training and testing [3]. Two studies [12, 118] have applied cross-validation, however, they have used a small dataset of 8 patients and only 130 minutes of EEG [118], and 7 patients and 121 minutes of EEG [12] respectively.

Furthermore, we reported several common metrics, i.e., the number of false positive detections, sensitivity and specificity to quantify the performance of IED detection. It is important to report the false detection rate especially when there are numerous IEDs in the EEG.

Moreover, the number of IEDs in the EEG recordings is not considered in many studies while evaluating the performance metrics. We observe from Fig. 3.11 that IED detection in EEGs that have numerous IEDs is generally easier than in records with few IEDs, as expected. One study by Casson *et al.* [135] suggests to include the frequency of IEDs in each record while calculating the performance, since the record with numerous

IEDs would skew the overall results. Therefore, in this study, we have reported the performance characteristics for individual EEG recordings based on their IED counts, as shown in Fig. 3.8 and Fig. 3.11.

3.5 Conclusions

In order to design an efficient system for IED detection, we proposed an algorithm for fast rejection of background data from scalp EEG of patients with suspected epilepsy. We considered various features including signal morphology, wavelet coefficients, nonlinear energy operator, and spectrogram features. By applying thresholds on these features in multiple steps, we formed a cascade and eliminated background waveforms in each stage. We showed that by applying this cascade, we are able to reject a large proportion of background waveforms, while maintaining most of the IEDs.

Since speed is a crucial factor for any IED detection system, we compared two classification scenarios. First we applied the classifiers including support vector machine, K-nearest neighbor, and random forests to classify EEG; secondly, we applied them after rejecting the background from EEG. The comparison between the performance and speed of these two methods revealed that background rejection speeds up IED detection by eliminating most irrelevant segments from the EEG at an initial phase. By utilizing this cascade method, we can save time while applying more complex pattern recognition methods for detecting IEDs in large EEG datasets.

We investigated a large dataset consisting of 156 EEG recordings including epileptic patients as well as healthy subjects. In addition, we applied cross-validation on the dataset. The large dataset in this study and the procedure ensure that this method can generalize to EEG recordings from other datasets.

Moreover, the feature ranking yielded in each stage reveals the most discriminative features. Therefore, we can employ these features to separate IEDs from background EEG activity and utilize them for IED detection.

One limitation of this method is that it does not improve the performance of IED detection, rather, it only lowers the computation time. While this is the main goal of developing this algorithm, it is meaningful to make modifications in such a way that we

can improve the results in terms of sensitivity and false positive rate as well.

To this end, we may apply alternative feature selection methods. So far, we utilized the CDF value to select features for each step. Exploring other methods might change the feature order, and possibly improve the results. Using quicker feature selection methods can also help us decrease the time needed to build the cascade.

Furthermore, we can increase the threshold, such that fewer IEDs would be lost in each stage. By setting higher thresholds and leveraging on more diverse EEG features, we may increase the length of the cascade, and achieve higher sensitivity and specificity. This requires an even larger dataset of IED and background EEG data to avoid overfitting. In addition, applying the method on an even larger number of subjects will provide more reliable performance characteristics of this method in general.

Our ultimate aim is to develop an efficient and reliable IED detector. Therefore, we will further process the data that remains after the background rejection algorithm by more complicated classification algorithms, which are usually more computationally demanding. By filtering out a large proportion of background data by our proposed algorithm, we hope to maintain a reasonable computational complexity for the overall IED detection system.

Chapter 4

Classifier Cascade to Aid in Detection of Epileptiform Transients in Interictal EEG

In this chapter, we propose an ensemble method in order to improve the performance of IED detection in terms of precision and false positive rate. We describe this technique and investigate its performance on a large dataset.

4.1 Introduction

In this study, we analyze a large dataset consisting of 156 interictal scalp EEG recordings. We propose a classifier cascade, i.e., an ensemble of weak classifiers to aid in classifying the interictal EEG. We establish this algorithm to exclude as many background EEG waveforms as possible from the EEG. We design each weak classifier is to have a high sensitivity to detect IEDs while allowing for modest or weak specificity. We apply these classifiers as a cascade; the cascade can then be combined with other IED detection algorithms to improve their performance.

4.2 Materials and Methods

In this section, we first describe the scalp EEG dataset considered in this study. Next, we present the proposed algorithm to help in IED detection. Afterwards, we explain the training and testing phases. Finally, we elaborate the approach to evaluate the performance of the algorithm.

In the clinical application of IED detection, there is limited tolerance for false alarms (or false positive rate), so that no healthy subjects will be diagnosed as having epilepsy. On the other hand, the true positive rate (sensitivity) does not need to be as high, since even with slightly lower sensitivity we can still detect IEDs in sufficiently long EEG recordings. The main objective in IED detection is to determine whether any IEDs exist in a patient’s EEG, and if so to find their channel locations [148]. Therefore, the aim of this study is to develop an algorithm to lower the false positive rate and increase the precision. Similar to Chapter 3, we consider a large EEG dataset in this study, in contrast to several existing studies. Applying cross-validation ensures the results generalize well.

In this work, we propose another ensemble method to aid in IED detection. We develop an algorithm in which several simple and quick classifiers compete in several stages, and the winner of the competition in each stage does one part of the classification task. We employ a support vector machine (SVM) in each stage. As explained in Chapter 3 (Section 3.2.3.3) SVM has several advantages, including the fact that it can be quickly trained and is not too susceptible to overfitting. It also performs well for high-dimensional data [145].

4.2.1 EEG dataset

We consider recordings of 93 patients with epilepsy, and 63 subjects with non-epileptic EEG. The diagnosis was based on the clinical read which was done independent of this study. Average length of each EEG is 28.5 minutes. Most previous studies have only considered the EEG from patients with epilepsy containing IEDs. By contrast, we included spike-free EEGs for both training and validation. This is beneficial since we need the algorithm to perform well on all types of patients. The data was acquired at Massachusetts General Hospital (MGH) using the 10-20 standard system of electrode

placement. The sampling frequency is 128Hz, and signal is high-pass filtered at 1 Hz. Data was cross-annotated by two trained clinical neurophysiologists using the Neuro-Browser software [149]. We only considered the waveforms annotated by both experts as IEDs. There are a total number of 18,164 IEDs in the dataset, with 143 IEDs per patient on average, and minimum 1 and maximum 1987 IEDs per patient. We applied 5-fold cross-validation to split the data into training and test sets. The EEG recording from each subject is randomly assigned to a fold, and the data from each subject is only utilized either for training or testing for each fold. We divided the EEG into segments of 0.5s, and casted the IED detection as a binary classification problem for each segment.

4.2.2 Classifier cascade

We developed a model ensemble by considering different single classification models. In ensemble learning, rather than creating a single model, a set of models are produced, and predictions are made by aggregating the outputs of these models. A prediction model which consists of a set of models is called a model ensemble. The motivation behind using ensemble methods is the idea that a committee of experts working together on a problem is more likely to solve it successfully than a single expert working alone [145].

In the proposed method, we generate several classification models having a high sensitivity. We compare these models in terms of their specificity and select the best one in each stage of the algorithm. The training set shrinks at each stage. We apply the selected classifiers at every stage on the dataset in the form of a cascade. The reason for reducing the training set at each stage is that the goal is to retain IEDs, while rejecting the majority of interictal EEG which consists of other EEG waveforms collectively called as background. Therefore, we narrow down the IED candidates at every stage of the algorithm. Moreover, as mentioned earlier, one obstacle in IED detection is low speed when applying detection methods to large datasets. Hence, to increase speed, it is beneficial to eliminate low-probability background waveforms at an early stage. We employ SVM [145] with Gaussian radial basis kernel as the basic unit of the classification cascade. Our hypothesis is that by sequentially eliminating background waveforms using SVMs applied as a cascade, while retaining possible IEDs, we can increase the precision and reduce the number of false positive detections of an IED detection method.

4.2.2.1 Training phase

We consider IEDs as 0.5s segments at the specific EEG time and channel annotated by experts. We eliminate all waveforms overlapping in time with any IED, and 0.5s before and after any IED, on all channels from background candidates. First, we split the data in 5 cross-validation folds. For each training round, we sample one background waveform for each IED, so that the training set is balanced. The average number of IEDs is 14,531 per fold. We train classifiers using each background set (from each EEG in training set) and all IEDs. For each fold, the number of trained SVMs equals to the number of training EEGs. Thus, we train around 124-125 SVMs for each fold, and 624 SVMs over all folds.

We then select one of the trained SVMs for the first stage, as follows: We first apply all trained SVMs on the whole training set, including all IEDs and the sampled background from all training subjects. Next, we adjust the threshold on the output scores of the classifiers in such a way that sensitivity is greater or equal to 0.999. We select the classifier having the highest specificity and apply it on the whole training set.

Hence, all waveforms in the training set which were labeled “non-ET” after applying the selected SVM, are eliminated from the dataset. In the next stage, again we apply all classifiers on the new training waveforms labeled “ET” in the previous stage and repeat the same procedure. We can determine the optimum number of stages based on desired performance metrics.

One purpose for training several classifiers with different training sets is to avoid unbalanced training. Since in interictal EEG, there are only few IEDs typically, the training set would be highly unbalanced if using one single classifier. In the proposed approach, each SVM is trained with a balanced input.

Moreover, we train each classifier using one EEG’s background data and all the IEDs. This ensures that every possible background morphology has been taken into account. Therefore, we apply the classifiers which generalize better on the whole data that includes all background variations. By using different classifiers at every stage of the cascade, we make use of various classifiers trained on different data, which makes the algorithm perform better on new test data.

4.2.2.2 Testing phase

In the testing phase, we apply the classifier cascade to all the EEG waveforms in the test set. In other words, in this phase we extract all IEDs, and consider background waveforms in form of 0.5 s segments with 50% overlap. By contrast, in the training phase, one background waveform was randomly selected for each IED, leading to a balanced training set. The test set contains an average of 3633 IEDs per fold. We calculate the sensitivity (recall, or true positive rate), and specificity (true negative rate) of the cascade after each stage.

4.2.3 Combining the cascade with an IED detection method

We test a single SVM as a benchmark and compare it to the classifier cascade followed by the same SVM. In this way, we can assess the benefits of the classifier cascade and see whether the proposed algorithm improves the performance. We illustrate these two approaches for IED detection in Fig. 4.1. After applying the IED detection method, we calculate and compare the main performance characteristics of the overall system. The performance characteristics are defined as:

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \quad (4.1)$$

$$\text{Specificity} = \frac{\text{True negatives}}{\text{True negatives} + \text{False positives}} \quad (4.2)$$

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \quad (4.3)$$

$$\text{False positive rate} = \frac{\text{False Positives}}{\text{EEG recording time (min)}} \quad (4.4)$$

We present and compare these values in the next section. The comparison helps determine whether using the proposed cascade improves IED detection performance.

Figure 4.1: Two approaches for IED detection: (a) One single SVM detector (b) SVM cascade followed by one SVM detector.

4.3 Results and Discussion

In this section, we present the results of applying the classifier cascade on the cross-validated EEG dataset. We demonstrate several performance characteristics and discuss how the results show promising improvements in IED detection performance.

- Classifier cascade

We determined the classifiers and thresholds to be applied on their output by keeping the sensitivity at 0.999 for every stage. After learning the classifiers, we applied them on the test set for every fold. The testing sensitivity and specificity of the algorithm for every fold as well as the average value over all 5 folds are listed in Table 4.1 and Table 4.2, respectively. After 10 stages, we achieved the specificity of 0.919 at sensitivity of 0.883. The overall sensitivity decreases as we add more stages, while the overall specificity increases, as expected. The choice of number of stages depends on the application.

The ratio of retained IEDs (sensitivity) and excluded background waveforms (specificity) for individual subjects are illustrated in Fig. 4.2. For the majority of subjects, the sensitivity and specificity values are high. However, for a few subjects the algorithm has low performance. The subjects with low sensitivity typically have very little number of IEDs (1, 4, 14), while one subject with sensitivity of 0.26 has 226 IEDs.

- IED detection by SVM

Two approaches were taken for IED detection: (1) train and test an IED detection method using a single SVM; (2) train and test an SVM on the data retained from the classifier

Table 4.1: Sensitivity of the classifier cascade after each stage.

Stage	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
1	0.999	0.982	1	1	1	0.996
2	0.996	0.971	0.998	0.998	0.908	0.974
3	0.973	0.962	0.991	0.986	0.908	0.964
4	0.970	0.962	0.991	0.985	0.852	0.952
5	0.962	0.849	0.991	0.970	0.848	0.924
6	0.959	0.847	0.991	0.970	0.824	0.918
7	0.930	0.847	0.984	0.962	0.795	0.904
8	0.907	0.843	0.984	0.953	0.794	0.896
9	0.904	0.842	0.952	0.953	0.789	0.888
10	0.904	0.841	0.952	0.9487	0.771	0.883

Table 4.2: Specificity of the classifier cascade after each stage.

Stage	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
1	0.533	0.784	0.323	0.577	0.586	0.561
2	0.738	0.848	0.408	0.730	0.754	0.696
3	0.836	0.888	0.556	0.828	0.792	0.780
4	0.851	0.908	0.668	0.861	0.843	0.826
5	0.875	0.915	0.720	0.886	0.866	0.852
6	0.887	0.927	0.762	0.904	0.879	0.872
7	0.903	0.933	0.792	0.918	0.896	0.888
8	0.910	0.940	0.825	0.927	0.915	0.904
9	0.914	0.945	0.828	0.934	0.923	0.909
10	0.925	0.950	0.851	0.941	0.930	0.919

Figure 4.2: Sensitivity versus specificity for each subject after a 10-stage classifier cascade.

cascade, considering a 4-stage cascade which includes 4 SVMs.

The reason for choosing 4-stage cascade is maintaining the balance between the desired sensitivity and precision.

We computed the sensitivity, precision, and false positive rate (FPR) per minute of EEG recording. The results for the overall algorithm are listed in Table 4.3. The performance is computed and compared on parts of ROC where the overall specificity of IED detection equals to 0.99, 0.999, and 0.9999. High values of specificity are considered, since we require the overall specificity of the system to be very high in order to control the false positives.

As shown in Table 4.3, sensitivity declines by applying the classifier cascade to eliminate background EEG before applying the IED detection method. This is expected, since at every stage of the cascade we lose a small portion of IEDs. As mentioned earlier, we are more interested in high precision than high sensitivity. Hence, we can compromise on sensitivity to some extent. On the other hand, we observe that precision is significantly improved, and false positive rate per minute is reduced by adding stages to the cascade.

The precision versus sensitivity of the SVM detector, with and without the cas-

cade, is shown in Fig. 4.3. This figure shows the precision versus sensitivity (recall) for the specific values of specificity=0.99, 0.999, 0.9999 similar to the values included in Table 4.3. The sensitivity versus false positive rate per minute on a logarithmic scale (with base 10), corresponding to each of the three mentioned specificity values, is shown in Fig. 4.4. From Fig. 4.3 we observe that the precision-recall curve for IED detection

Table 4.3: Overall performance of IED detection by applying SVM with and without the initial classifier cascade.

Method	Specificity=0.99			Specificity=0.999			Specificity=0.9999		
	Sensitivity	Precision	FPR	Sensitivity	Precision	FPR	Sensitivity	Precision	FPR
SVM	0.806	0.078	42.76	0.577	0.369	4.289	0.325	0.715	0.432
1stage+SVM	0.803	0.088	43.041	0.602	0.405	4.277	0.347	0.739	0.430
2 stages+SVM	0.689	0.284	17.254	0.501	0.572	1.705	0.242	0.819	0.163
3 stage+SVM	0.673	0.281	9.823	0.472	0.615	1.232	0.218	0.842	0.116
4 stage+SVM	0.634	0.346	6.635	0.425	0.661	0.828	0.186	0.865	0.079

with the classifier cascade is shifted upwards compared to the case in which only a SVM detector is applied to the whole dataset. This shift in the precision-recall curve indicates that by including the cascade of classifiers as the first stage, we can achieve higher precision for the same sensitivity. Fig. 4.4 shows that the false positive rate is significantly reduced by including the cascade before applying SVM IED detector, and by increasing the number of classifiers included in the cascade.

According to Fig. 4.4, for a fixed false positive rate of 1.2 per minute (-0.081 in logarithmic scale), sensitivity is increased from 39.70% to 42.53% which indicates a total increase of 2.83% in sensitivity.

Fig. 4.3 shows that for a fixed sensitivity of 42.53%, precision is increased from 59.38% to 66.16%, indicating a total increase of 6.78% in precision.

In summary, the results show that the proposed algorithm can successfully increase the precision and reduce the number of false positive detections of an IED detection system. The classifier cascade, which is an example of ensemble methods, can

Figure 4.3: Precision versus sensitivity at certain specificity values, with and without the classifier cascade.

Figure 4.4: False positive rate versus sensitivity at certain specificity values, with and without classifier cascade.

be incorporated with any IED detection system as an initial step. It not only improves the performance, but also saves computation time by eliminating large amounts of back-

ground data at the beginning of the IED detection process.

4.4 Conclusions

In this study, we proposed a method to reject background EEG activity from interictal EEG of patients with epilepsy by applying an ensemble of classifiers as a cascade. The purpose is to improve the performance of IED detection systems in terms of precision and false positive rate. We considered a large dataset consisting of 93 epileptic subjects as well as 63 subjects with IED-free EEG. We employed cross-validation for training and testing as well. This approach ensures the reliability of the algorithm for clinical applications on new EEG data which might or might not be epileptic. We applied three IED detection methods and compared their performance with and without the SVM cascade as a benchmark. The results show that by applying the SVM cascade, at a fixed sensitivity the IED detection precision improves by 6.78%; and at a fixed false positive rate the sensitivity improves by 2.83%.

In the future work, we hope to apply this method on an even larger dataset to provide more reliable performance. In addition, other classifiers as basic units of the system can be employed, in order to increase the background rejection rate and further improve precision and false positive rate. Various sizes for cascade can be considered as well, to pick the optimum for our application. We can also investigate the algorithm with more IED detection methods.

Furthermore, more analysis can be performed on individual subjects to determine why the algorithm does not perform very well for certain patients. Understanding the characteristics of the patients on which our algorithm fails to perform well could give clues to tune the IED detection algorithms to handle these difficult cases.

Finally, we can analyze IEDs that are lost at the early stages of the cascade. In this way, we can learn more about morphologies of IEDs that can easily be misinterpreted by IED detection algorithms. We can also analyze the background waveforms which are retained after the late stages of the algorithm. These waveforms comprise difficult-to-reject false positive detections; therefore, studying their morphologies and trying to eliminate them could yield further reduction in the false positive rate of IED detection

methods in the future.

Chapter 5

Interictal Epileptiform Discharge Characteristics Underlying Expert Interrater Agreement

In this chapter, we analyze the inter-rater agreement among experts in interpreting interictal EEG; and we present several methods to quantify this agreement and to predict the expert opinion. Additionally, we investigate the features and characteristics of EEG leading to expert agreement.

5.1 Introduction

We described in Chapter 2 that there is a substantial discrepancy in expert interpretations of interictal EEG. The factors underlying agreement and disagreement regarding IEDs among human experts has not been intensively studied. IED-detection software could perhaps be improved if we understood more which features optimally characterize IEDs, and which features lead to disagreement between experts. In this study, we test a broad list of EEG features for their potential correlation with expert scored IEDs, test different approaches to EEG normalization before feature extraction for the purpose of predicting expert scores, and explore which features contribute to inter-rater agreement (IRA) and disagreement among experts.

5.2 Materials and Methods

In this section, we explain the methods utilized in the present study for data collection, signal analysis, inter-rater agreement analysis, and machine learning techniques to predict expert scoring.

5.2.1 EEG scoring and EEGnet software

EEG recordings were annotated by experts using EEGnet, a software system hosted at the School of Computing of Clemson University. By means of this software, users can view and annotate EEG datasets on a web browser with no local software installation required [7]. EEGnet has been designed to create a user interface similar to commercially available software applied in clinical EEG interpretation. Ten seconds of EEG data are displayed at one time; it also shows channel labels along with one electrocardiogram channel. Channel gain can be adjusted for all channels or for individual channels, and users can scroll through the signal one or ten seconds at one time. Several conventional EEG montages which include: hatband montage, transverse bipolar, AP bipolar, Cz reference, ipsilateral ear reference, and common average reference can be considered for scoring [7].

5.2.2 EEG dataset description

A dataset of EEG segments has been formed by reviewing around 1000 EEG recordings performed during the years 2010-2011, at the Medical University of South Carolina (MUSC) Neurophysiology Laboratory, for clinical purposes. EEGs were recorded by applying the standard 10-20 electrode placement. 200 thirty-second EEG segments from EEG recordings of 200 different patients were chosen by two experts. Fifty segments were selected from EEGs which contained IEDs. Fifty segments were from EEGs read as normal in the clinical report. Fifty segments were chosen due to containing IEDs from patients who were known to have epilepsy with difficult-to-interpret IEDs. Fifty were chosen because of containing benign paroxysmal activity; an inexperienced reviewer might misinterpret these segments to be epileptiform. Institutional Review Board (IRB) approval which allowed collection and retrospective review of de-identified clinical EEG

data as well as sharing the dataset with researchers at other universities has been granted by MUSC and Clemson University. All 200 segments were randomly ordered and concatenated into one long 100-minute recording. All EEG records are high-pass filtered by 0.5 Hz and resampled to 256 Hz. Notch filter at 50Hz is applied on all records as well. In addition, we remontage all records to common average montage.

5.2.3 Scoring the EEG dataset by experts

Annotation of IEDs by experts was conducted in three phases. During the first phase, 19 academic clinical neurophysiologists (EEGers) were instructed to label all IEDs on the channel in which they thought it was best represented. All of these experts had completed at least one year of fellowship training in EEG interpretation and were board certified by the American Board of Clinical Neurophysiology. A clustering algorithm was applied to eliminate redundancy by merging labeled EEG segments that strongly overlap in time, as described in previous studies [7]. If two labeled EEG segments overlapped in time by at least half of the duration of the shortest segment, they were placed into the same cluster. The representative EEG segment for a cluster was considered as the EEG segment which had maximum sum of correlations with all segments of the same cluster. A master list of EEG events was formed by adding all the representative EEG segments, to be considered in the second phase of scoring. The master list included a representative segment from each cluster which included at least two members (events which were marked by at least two EEGers.)

Of the 19 EEGers who participated in phase one of the study, 18 EEGers took part in the second phase of EEG scoring. In this second phase, each EEGer was presented with the same EEG recording in which the master list of 235 events was labeled within it, presented as transparent yellow boxes superimposed over single channel EEG segments. The EEGers were guided to click on each highlighted segment and to score it on a 5-point Likert scale (using keyboard keys 1-5) as either (1) 'definitely not an epileptiform discharge', (2) 'not likely an epileptiform discharge', (3) 'not sure, could go either way', (4) - 'likely an epileptiform discharge', and (5) 'definitely an epileptiform discharge'.

In some cases, EEGers were inconsistent in their marking of IEDs between phase

one and phase two scoring. For example, in some cases an EEGer marked an event as an IED in phase one scoring yet in phase two scoring labeled the same event with a score of 1-3, indicating that they thought the event was probably not an IED. In the third phase of scoring, each EEGer was presented with a unique list of events for which they were inconsistent between phase one and phase two scoring. Each EEGer was then asked to rescore this list of events in EEGnet using the phase two scoring method. This provided a third “tie breaker” opinion on these events. Further details about this method are published elsewhere [150].

5.2.4 Inter-rater agreement analysis

Here we explain two inter-rater agreement coefficients which we can apply to analyze the expert scores: (1) Fleiss’ kappa ,and (2) linearly weighted Gwet’s agreement coefficient (AC2).

1. Fleiss’ kappa [151]

Fleiss’ kappa is a statistical measure for assessing the reliability of agreement. For this test, each sample of subjects is rated on a nominal scale by the same number of raters. It can be applied for any number of raters giving categorical ratings to a fixed number of items. It expresses the extent to which the observed degree of agreement among raters exceeds the expected agreement if all raters made the ratings randomly. We can consider N as the total number of subjects, n as the number of ratings per subject, and k as the number of categories. The subscript i , where $i = 1, \dots, N$, shows the subjects, and the subscript j , where $j = 1, \dots, k$, shows the scale categories. n_{ij} is defined as the number of raters who assigned the i th subject to the j th category, and

$$P_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij},$$

$$1 = \frac{1}{n} \sum_{j=1}^k n_{ij}.$$
(5.1)

The degree of agreement among n raters for the i th subject can be indexed by the ratio

of agreeing pairs out of all $n(n - 1)$ possible assignment pairs. This ratio is:

$$\begin{aligned} P_i &= \frac{1}{n(n-1)} \sum_{j=1}^k n_{ij}(n_{ij} - 1) \\ &= \frac{1}{n(n-1)} \left(\sum_{j=1}^k n_{ij}^2 - n \right). \end{aligned} \quad (5.2)$$

The overall degree of agreement may then be quantified by the average of P_i s:

$$\begin{aligned} \bar{P} &= \frac{1}{N} \sum_{i=1}^N P_i \\ &= \frac{1}{Nn(n-1)} \left(\sum_{i=1}^N \sum_{j=1}^k n_{ij}^2 - Nn \right). \end{aligned} \quad (5.3)$$

Some extent of agreement is expected solely by chance; if the raters made their assignments completely randomly, the mean proportion of agreement is expected to be:

$$\bar{P}_e = \sum_{j=1}^k p_j^2. \quad (5.4)$$

The degree of agreement obtainable above what would be predicted by chance is quantified by $1 - \bar{P}_e$. The degree of agreement actually obtained above chance is $\bar{P} - \bar{P}_e$. Therefore, a normalized measure of overall agreement, corrected for the agreement expected by chance, would be:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}. \quad (5.5)$$

$\kappa = 1$ in the case that the raters fully agree with each other, while $\kappa \leq 0$ if there is no agreement other than what is expected by chance. However, there are some short-comings for this test [152, 153]:

- Kappa is not a chance-corrected measure of agreement;
- Kappa can be low even when there is high degree of agreement and even when individual ratings are accurate. Whether a Kappa value indicates a good or a bad rating system or diagnostic method depends on the model assumed about the decision making of the raters [154];

- Kappa does not distinguish among different types and sources of disagreement;
- Kappa is affected by trait prevalence (distribution) and base rates. Consequently, they are seldom comparable across studies, procedures, and populations [155–157].

2. Linearly weighted Gwet’s agreement coefficient (AC2) [153, 158]

This measure can be adjusted for both chance agreement and misclassification errors, that Kappa coefficients do not allow.

The second-order agreement coefficient, or AC^2 statistic, can be employed for any number of raters using an ordered categorical rating system to rate objects.

Using an ordered categorical scale, some disagreements are considered more serious than others. If two raters assess an item similarly, but classify the items differently, it is considered as a classification error. If two categories are similar, the probability that an independent rater would classify the patient initially classified into one category, into the other category based on the initial reviewer’s assessment would be high. But if the categories are quite different, the probability for such classification would be low.

Misclassification errors can cause bias in the inter-rater reliability estimate. The second-order agreement coefficient can be adjusted for chance-agreement and misclassification errors, if the misclassification probabilities are given.

In order to adjust for chance-agreement, one can get an estimate of the overall probability of agreement and condition it on being no chance agreement. The overall probability of agreement is represented as P_a , and chance agreement is represented as P_e .

P_e is the probability that raters have agreement, and the rating is made by chance. Hence, the chance-adjusted agreement is:

$$AC = \frac{P_a - P_e}{1 - P_e}. \quad (5.6)$$

Marginal homogeneity was assessed using the Stuart-Maxwell test [159, 160]. To quantify agreement among the experts, we measured the IRA using linearly weighted Gwet’s agreement coefficient (AC2) [161, 162]. We did not consider Kappa measures of IRA due to the “kappa paradox” which leads to lower IRA values for datasets with low marginal

homogeneity [163].

We calculated Gwet's AC2 for all possible subsets of the 18 scorers to give us the measure of agreement between all scorers, and to find a subset of scorers of a desired size having the highest inter-rater agreement compared to the other ones.

In order to interpret the agreement, AC2 can be interpreted as: 'very good' 0.8-1.0, 'Good' 0.6-0.8, 'Moderate' 0.4-0.6, 'Fair' 0.2-0.4, 'Poor' less than 0.2. We selected groups of EEGers of size 2 to 17 having highest AC2 among themselves as groups of "top scorers". We considered the majority opinion of all EEGers and the top scorers' groups for training machine learning models, as we discuss in the following sections.

5.2.5 Signal analysis methods

We explain the approaches taken for preprocessing of EEG signals and feature extraction as follows.

5.2.5.1 EEG signal preparation

EEG records were resampled at the sampling frequency of 256 Hz, and high-pass filtered at 0.5 Hz. Since various events were annotated on one of seven montages, they were all re-montaged to common average referential montage in order to make the analysis uniform. Although the events were annotated with differing lengths, we analyzed them all as 0.5 s segments centered at the peak of the IED.

5.2.5.2 Feature extraction

In order to determine which signal features are more helpful in predicting the rater agreement or disagreement, we extract a broad range of EEG features from the annotated events; we then apply correlation analysis and feature selection to determine the most relevant features. We also employ these features to visualize the events in 2D using the t-SNE algorithm. We analyze the relationship between expert scores and the electrode locations of each event. Finally, using the selected features, we fit a regression model to the data to predict the average score given to each event by the experts.

All data processing and statistical analysis were performed off-line using a com-

mercial software package (MATLAB 2014a with the Statistical Analysis Toolbox, The Math Works Inc., Natick, MA, 2014). Many types of signal analysis features are evaluated: (1) morphological features, (2) nonlinear energy operator (NLEO), (3) wavelet analysis, (4) energy of the EEG signal, (5) power spectral density using spectrogram, and (6) scalp electrode location. Methods for feature extraction are explained below. A total of 5538 features were computed for the EEG signals.

5.2.5.3 Signal morphology features

These features are related to the shape and morphology of the EEG waveform. The considered features include: (a) peak voltage, (b) rising voltage and slope, (c) falling voltage and slope, (d) peak-to-peak voltage, (e) spike duration, and (f) line length. We performed feature extraction on the signals which were bandpass-filtered in the frequency interval of 0.5-256 to 128 Hz, as well as 5 main EEG frequency bands and their combinations. The frequency band is divided as: <4 Hz, 4-8 Hz, 8-12 Hz, 12-32 Hz, >32 Hz, which are known as delta, Theta, Alpha, Beta, and Gamma bands, respectively [164]. The definitions of morphological features are illustrated in Fig. 5.1. First, we found the IED peak

Figure 5.1: Morphological features of an interictal epileptiform discharge.

in each waveform. The neighboring troughs were then computed, and the desired voltages and slopes as well as the spike duration were calculated. We defined peak to peak

voltage as the difference between the highest peak and lowest trough in the segment. We calculated the slopes based on the rising and falling edge voltages, and we defined the duration as the total time in which the voltage goes up to the spike peak and falls down to the trough. The line length of signal $x[n]$ with N samples was computed as [139]:

$$L = \sum_{n=1}^N |(x[n-1] - x[n])|. \quad (5.7)$$

5.2.5.4 Nonlinear energy operator

NLEO is useful for IED detection purposes. We explained this feature in Chapter 2 (Section 3.2.2.4). The resolution parameter k [143] is chosen to have values in the range of 1 to 40. First, we calculated the NLEO value for every sample point of the EEG signal before extracting the event segment. Then we considered the NLEO features for each 0.5s waveform as the standard deviation, maximum, minimum, and maximum absolute NLEO value within the segment. `segment.subsec:BG:wavelets`

5.2.5.5 Wavelet analysis

The wavelet transform is helpful for characterizing several aspects of EEG signals such as non-stationary transient events which makes it suitable for IED detection purpose [3]. We explained this feature in Chapter 2 (Section 2.2.5) and Chapter 3 (Section 3.2.2.5). In this study, we applied the discrete wavelet transform (DWT) to decompose the signal into multiple scales based on wavelet basis functions. Since there is no consensus as to which wavelet basis function is best suited to characterize IEDs, we applied multiple wavelet basis functions to extract wavelet features. Hence, in this study we considered 53 wavelets including symlets (sym), coiflets (coif), Daubechies (db), biorthogonal (bior), reverse biorthogonal (rbio), and discrete approximation of Meyer (dmey) wavelet basis families with different degrees. 5 levels were considered for applying the DWT, and the wavelet decomposition was obtained for all levels [165]. The detail and approximation coefficients were then reconstructed for every level. The output of this step is a vector of same length as the input segment. We calculated the features based on the resulting detail and approximation coefficients; these features include standard deviation, maximum,

minimum, and maximum absolute value of the detail and approximation coefficients at every level, within each segment for the transforms performed using each of the wavelet basis functions. Since the sampling rate is 256 Hz, the filters applied in the wavelet transform for each of the decomposition scales include the frequency bands of d1 (64 – 128 Hz), d2 (32 – 64 Hz), d3 (16 – 32 Hz), d4(8 – 16 Hz), d5 (4 – 8 Hz), a1 (0.5 – 64 Hz), a2 (0.5 – 32 Hz), a3 (0.5 – 16 Hz), a4 (0.5 – 8 Hz), and a5 (0.5 – 4 Hz).

One potential disadvantage of the critically sampled discrete wavelet transform (DWT) for a one-dimensional signal is the lack of shift invariance; i.e., a small amount of shift in the signal can significantly perturb the wavelet coefficients. Since we need our method to generalize well for new data and signals which are shifted with respect to their peaks, we further investigated whether this can be problematic for this study or not. Complex wavelets provide a solution to this shortcoming of the DWT as they are known to be shift invariant. Therefore, in addition to the DWT features described above, we also evaluated the dual-tree complex wavelet transform (DT-CWT), in which two separate DWT decompositions are calculated providing real and imaginary coefficients, and using specifically designed filters [166, 167]. This method is applied here to decompose the signal into multiple scales in multiple levels. The filters ‘dtf1’, ‘dtf2’, ‘dtf3’, ‘dtf4’ were applied for the wavelet decomposition. Features were selected based on the resulting detail and approximation coefficients similar to the features extracted from DWT.

Because we extract summary statistics of DWT coefficients as features (e.g. maximum value within a brief window) rather than using the wavelet coefficients directly, it is possible that the effect of shifts will be negligible for this study. Therefore, we applied the DWT on the shifted signals in order to find out whether shifts adversely affect the prediction of expert scores. By evaluating the DT-CWT in addition to DWT on shifted signals, we determine which type of wavelet transform better suits this analysis and objectives.

5.2.5.6 Energy of the signal

We computed the energy of the EEG signal for all possible frequency bands spanning at least 4 Hz, starting at 1 Hz in steps of 2 Hz. We first filtered the signal by a linear phase FIR filter which was applied backward and forward to ensure zero-phase filtering,

and then computed the energy. We calculated the energy of a discrete signal $x[n]$ with n samples as

$$E = \sum_{n=1}^N x[n]^2. \quad (5.8)$$

5.2.5.7 Spectrogram and power spectral density

We derived the spectrogram of the signal by applying the short time Fourier Transform (STFT) with Hamming windowing [65]. We defined several STFT features based on the matrix which consists of PSD values for each time segment of the signal, as follows: (a) mean of the PSD values, (2) maximum PSD value, (3) frequency and time of the maximum PSD value, (4) mean PSD value in neighborhoods of the maximum of the PSD in both time and frequency axes, considering all possible symmetric neighborhood lengths, (5) standard deviation of the mean PSD values along both frequency and time axes, as well as (6) mean of the standard deviation of these values.

We carried out the STFT and PSD derivation from it, as follows:

STFT in practice is computed by dividing the signal into short segments of equal length and computing the Fourier transform on each segment separately. Changing of spectra is usually plotted as a function of time.

In discrete time, the data can be divided into frames which usually overlap to reduce boundary artifacts. After deriving the Fourier transform of each frame, the complex result is added to a matrix recording magnitude and phase for every point in frequency and time. Here we provide the derivation of STFT and PSD [65, 168]. Having signal $x[n]$ and window $\omega[n]$, we can express it as:

$$STFTx[n](m, \omega) \equiv X(m, \omega) = \sum_{n=-\infty}^{n=\infty} x[n]\omega[n-m]e^{-j\omega n}. \quad (5.9)$$

Since STFT is mostly computed using Fast Fourier Transform (FFT), both m and ω are discrete values.

The spectrogram is yielded by the magnitude squared of the STFT:

$$spectrogram(x[t])(\tau, \omega) \equiv |X(\tau, \omega)|^2. \quad (5.10)$$

However, in this study we consider one-sided modified periodogram estimate of the PSD of each segment. The elements of the PSD matrix P are given by:

$$P(i, j) = k|S(i, j)|^2, \quad (5.11)$$

where S is the STFT of the input signal, and k is a real-valued scalar defined as below for the one-sided PSD:

$$k = \frac{2}{F_s \sum_{n=1}^L |w(n)|^2}, \quad (5.12)$$

where $w(n)$ denotes the window function (Hamming window chosen in this study), and F_s is the sampling frequency. We employed MATLAB Signal Processing Toolbox to compute PSD values.

5.2.5.8 Electrode location features

Our purpose is to find out which locations on the scalp are more probable to have events which are assigned to high scores by the experts.

We created several features to reflect the spatial location of IEDs on the scalp. Projecting the location of the EEG electrodes onto a two-dimensional scalp map, we assigned features to represent (1) anterior versus posterior location (2) temporal versus non-temporal location, (3) the lateralization (right versus left), and (4) the lateral distance from the midline (central versus lateral). A schematic of the 2D scalp locations is illustrated in Fig 5.2. We represented the 2D location of each channel by X (anterior-posterior) and Y (left-right) ranging from 1 to -1, respectively, based on the relative location of the electrode on the 2D scalp map. Temporal region versus non-temporal region is defined as $|X|$. Central versus non-central region is defined as $|Y|$.

5.2.5.9 Features versus expert scoring

We analyzed the correlation between the extracted features and two aspects of the EEGer scores. First, we analyzed the correlation between features and the average score of all experts ($\bar{s}$) to find which features are present in EEG events which the EEGers thought were IEDs. Second, we analyzed the correlation between features and the standard de-

Figure 5.2: The 2-dimensinal scalp EEG electrodes and the X and Y coordinates considered to represent the electrode positions.

viation of expert scores (σ_s) to determine which features were associated with the most disagreement among the EEGers.

In statistics, any statistical relationship between two random variables or two sets of data is known as dependence. Correlation refers to any statistical relationship involving dependence.

- Pearson correlation coefficient

The most common correlation coefficient is the Pearson correlation coefficient, which is sensitive to linear relationship between two variables. However, it may exist even if one variable is a non-linear function of the other. More robust correlation coefficients have been developed which are more sensitive to non-linear dependencies [169–171]. The most familiar measure of dependence between two quantities is the Pearson's correlation coefficient between two random variables X and Y denoted by $\rho_{X,Y}$:

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}, \quad (5.13)$$

where μ denotes the mean value, σ is standard deviation, cov is covariance, and E shows expected value. If we have a series of n measurements of X and Y written, the

sample correlation coefficient is utilized to estimate the population correlation. The sample correlation coefficient, r_{xy} is calculated as:

$$r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{n S_x S_y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}, \quad (5.14)$$

where $\bar{x}$ and s_x are sample mean and standard deviation, respectively.

- Spearman's correlation coefficient

Spearman's correlation coefficient is a rank correlation coefficient, and measure the degree to which by a increase in one variable, the other tends to increase; it is not required to follow a linear relationship. In case the increase in one variable, leads to decrease in the other, the rank correlation coefficients would be negative, similar to Pearson correlation. To be more precise, Spearman's rank correlation coefficient is considered as a non-parametric measure of statistical dependence and assesses how well the relationship between two variables can be expressed by a monotonic function. If there are no repeated values, Spearman correlation of +1 or -1 occurs when each variable is a perfect monotone function of the other [172, 173]. It is defined as the Pearson correlation coefficient between the ranked variables [174]. X_i and Y_i are converted to ranks rg_X , and rg_Y , then r is computed as:

$$r = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}. \quad (5.15)$$

In order to take into account any nonlinear relation between scores and features, we select Spearman's correlation to pick features that can be predictive of expert agreement.

5.2.5.10 Visualisation by t-SNE

T-SNE is a method that visualizes high dimensional data by assigning a location in a two- or three-dimensional map, to each data point. It is a helpful method especially for high dimensional data that lies on several different but related low dimensional manifolds [175]. Since the dimension of feature space is high in this study, we apply t-SNE to visualize the events according to the scores assigned to them by experts. This algorithm

reduces the dimensionality of the input data to 2 dimensions.

5.2.5.11 Feature selection: Elastic net regularization

Because of the large number of features which were generated in our analysis, we applied a feature selection step utilizing elastic net regression in order to select features which most significantly relate to the expert scores before predictive modeling analysis. Elastic net is a regularized regression method which is a hybrid of ridge regression and LASSO (least absolute shrinkage and selection operator) regression. LASSO is a method for performing linear regression which is a shrinkage estimator, i.e. it estimates the coefficients such that they are biased to be small and sparse (small number of nonzero coefficient values). The elastic net method can outperform LASSO and overcome its limitations, especially for data having highly correlated predictors and high-dimensional data with few examples [176, 177]. In order to avoid overfitting and to evaluate the performance of our method on test data, we applied a 10-fold cross-validated elastic net feature selection procedure to the feature sets. We performed the feature selection for each of the 10 cross-validation folds resulting in 10 selected feature sets which we then applied to train and evaluate the machine learning model that will be explained in Section 5.2.5.13. In addition, we applied a 10-fold cross-validation to select the parameter α , which determines the relative contributions of LASSO versus ridge optimization. We chose the value of α giving the least error and applied it to select the features to be included in the machine learning model.

5.2.5.12 EEG normalization

In order to determine whether normalizing the EEG data is a beneficial pre-processing step before feature extraction, we performed the analysis using multiple types of normalization procedures. We analyzed four different data representations based on four different normalization procedures: (1) no normalization; (2) global standardization of each EEG channel X by z-scores, performed as follows:

$$\bar{X} = \frac{1}{N} \sum_{n=1}^N x[n], \quad (5.16)$$

$$\sigma(X) = \sqrt{\frac{\sum_{n=1}^N (X[n] - \bar{X})^2}{N}}, \quad (5.17)$$

$$X_z = \frac{X[n] - \bar{X}}{\sigma(X)}, \quad (5.18)$$

where X represents each channel of the EEG signal with standardization performed individually for each channel in the EEG record by subtracting its mean and dividing by its standard deviation, and N represents the total number of samples in each EEG channel; (3) local standardization using a z-score performed for each data segment rather than for each channel, by applying (5.17) and (5.18) with X representing a 0.5s waveform, and N representing the number of samples in each waveform (128 points); and (4) RMS (root mean squared) normalization X_m calculated for each channel (globally) as:

$$RMS_{BR} = \sqrt{\frac{\sum_{n=1}^L (Br[n]^2)}{L}}, \quad (5.19)$$

$$X_m[n] = \frac{X[n]}{\text{median}[RMS_{BR}]}, \quad (5.20)$$

where $Br[n]$ represents the 0.5s background segments extracted by applying a sliding window for each channel of the EEG recording, and RMS_{BR} is the RMS of these segments. We define background as all the data in each 30 second EEG segment which is not annotated as an IED, and which does not overlap with the annotations in any of the channels. L is the number of samples in each background segment (128 points). The signal of each channel $X[n]$ was divided by the median of the RMS values of all background waveforms, yielding X_m .

5.2.5.13 Machine learning analysis

We trained machine learning regression models to predict the mean ($\bar{S}$) and standard deviation (σ_s) of the EEGer scores for groups of scorers of various sizes. These groups included all EEGers and subsets of EEGers of various sizes selected based on having the highest AC2 IRA coefficient for that group size. This allowed us to investigate how

well IED detection models performed with data from groups of EEGers of different sizes. Furthermore, we considered different representations of the data as inputs to fit the model including (a) the raw EEG data, (b) all extracted features, and (c) the features retained after feature selection with elastic net regression. We compute the root mean squared error (RMSE) and normalized root mean squared error (NRMSE) to assess the models. The RMSE is defined as:

$$RMSE = \sqrt{\text{mean}([X_p - X_e]^2)}, \quad (5.21)$$

where X_p and X_e represent variables that are predicted by the regression model and the actual value of the variable (mean or standard deviation of scores), respectively. The NRMSE is defined as:

$$NRMSE = \sqrt{\frac{\text{mean}([X_p - X_e]^2)}{\text{var}(X_e)}}, \quad (5.22)$$

If the NRMSE value is close to 0, there is a low error in prediction and the prediction is reliable, while the NRMSE value close to 1 suggests that the prediction is close to random guessing.

We trained support vector regression models to determine which combinations of features are present in the events with high ($\bar{S}$) or σ_s , and to predict the $\bar{S}$ and σ_s by these features. Support Vector Regression (SVR) is a variant of Support Vector Machine (SVM). SVM are a group of machine learning algorithms in which small subset of the training examples are used to represent the decision boundary between classes; these training examples are called the support vectors [178]. In this study, we applied a variant of SVM, ν -Support Vector Regression (ν -SVR) [179] with a radial basis function kernel to fit a regression model, which predicts the average expert score given to an EEG event. We applied the Library for Support Vector Machines (LIBSVM) [180] for this purpose. We applied 10-fold cross validation for training and testing the ν -SVR model. We tuned the cost parameter (c) by minimizing the cross-validation error for each fold. We trained SVR models on 9 folds and the trained model was applied on the 10th (remaining) fold, resulting in one predicted score for each event in the 10th fold. We performed this process over all 10 folds and computed the average RMSE and NRMSE over the 10 folds.

Throughout the analysis, including feature selection and regression, we applied internal cross-validation to select the model parameters. This helps tune model parameters while avoiding overfitting. In addition, since our dataset is relatively small, external cross-validation is important for evaluating the performance of the models and calculating error values on test data. Hence, we have applied nested cross-validation. In the usage of nested cross-validation (CV) for this study, we applied internal cross-validation to select values for α and then to train the model to obtain regression parameters giving least error on internal cross-validation. Then we applied external cross-validation to evaluate the performance of the regression model.

More specifically, in each round of external cross validation, the data is divided into training data and testing data; the testing data is also referred to as the “external validation data”. The training data is further divided into two sets: the majority is considered for model training, while the remaining data is utilized for internal CV. Internal CV consists of sweeping through a range of values for the model parameters (e.g. α) and evaluating model performance on the internal validation data. This process generates a curve of performance vs model parameters, which is used to select the optimal parameter value for the model, α . This parameter value α is then used to train the classifier model on all the training data. Note that, in the case of elastic net regression, choosing the regularization parameters determines which features are included in the classification model. The final classifier is then tested on the left-out testing data, yielding an estimate of out-of-sample performance. This entire “nested CV” procedure is repeated multiple times, rotating which subset of the data serves as the testing data. The final estimate of the out-of-sample testing performance is computed as the mean standard deviation of the performance across all folds of external cross validation.

5.3 Results and Discussion

In this section, we present the results of IRA, as well as expert score prediction; and we provide a discussion on the implications of the obtained results.

5.3.1 Expert annotation results

We investigate the degree of agreement among experts by calculating their relative sensitivity and specificity, as well as computing the two agreement coefficients for the scores.

The total number of events labeled in each category of the Likert scale for each scorer is shown in Fig. 5.3. As shown in this figure, there is no consistent pattern of scoring, and different scorers have different style of scoring. We calculate sensitivity and

Figure 5.3: Number of scores each expert has given to the events in total.

specificity for each EEGer versus the majority opinion of the other 17 EEGers. More precisely, this calculation is done with respect to the ground truth established by the remaining 17 EEGers. We consider two-group scores (IED, or non-IED) to define the ground truth.

It is worth mentioning that specificity is not usually calculated in IED detection studies because defining a true negative is difficult. However, by means of this unique two-phase approach to EEG annotation, we can define true negatives concretely. In this study, we define true negative detection as an EEG event, which was marked in phase one of scoring as an IED by two or more experts, but in phase two was annotated as non-IED

by the majority of experts and was labelled as non-epileptiform by an individual expert.

Sensitivity versus specificity for the 18 EEGers versus consensus EEGer opinion is shown in Fig. 5.4. We observe a high variance in the sensitivity and specificity values for each EEGer. Clearly, we can observe a significant difference between EEGers; some of them concern more about sensitivity, while the rest concern more about specificity.

Figure 5.4: Sensitivity and specificity of experts versus consensus of experts' opinion.

Overall, for categorizing epileptiform activity (rated 4-5 on scale) versus non-epileptiform activity (1-3 on scale), the inter-rater agreement was low to moderate in agreement with previous studies [45, 79, 134, 136]. Inter-rater agreement coefficient is 0.3249 for 2-level scores, and 0.1356 for 5-level scores.

In summary, we observe that: (1) there is a low level of agreement between experts in interictal EEG interoperation; (2) each EEGer has a pattern of EEG analysis different from others; (3) we can find a group of EEGers with higher level of agreement among themselves, which could be probably beneficial to characterize IED features

which are agreed by those experts.

5.3.2 EEG waveforms

All 235 events which are scored by 18 experts are categorized into three groups based on their scores in ascending order. The Figures Fig. 5.5, Fig. 5.6 and Fig. 5.7 demonstrate the change in the morphology of waveforms which causes EEGers assign different scores to them. The waveforms have higher peak voltage and more rapid changes in the voltage as the score increase. We observe that generally the waveforms in Fig. 5.5 (low score events) have much lower amplitude compared to Fig. 5.6 (medium score events) and Fig. 5.7 (high score events); and the waveforms in Figure S3 (high score events) have considerably higher amplitude and higher frequency alterations compared to the two other figures. In order to quantify the morphological features of these events, we

Figure 5.5: Waveforms with score: 1.5-2.38 sorted from low to high average score. Each waveform lasts 0.6 s at the same amplitude scale.

show the values for peak voltage, rising edge voltage, falling edge voltage, and peak to

Figure 5.6: Waveforms with score: 2.38-3.3 sorted from low to high average score. Each waveform lasts 0.6 s at the same amplitude scale.

peak voltage for all events as well as the events having the average score of 4 and higher in Fig. 5.8 as box plots. In addition, the mean values of the same features are listed in Table 5.1. We also applied a two-sample t-test on the two groups. In this study, the null hypothesis (H) is that the feature values for all events and events scored ≥ 4 come from independent random samples of normal distributions having equal means and equal or unequal unknown variances. And the alternative hypothesis is that the data comes from populations with unequal means. After applying this test, the null hypothesis was rejected for all four features. The p-values are listed in Table 5.1. Therefore, there is a significant difference in the feature values of these two groups.

From Fig.5.8, we observe that the median and corresponding quartile values are higher in the events with higher scores. Table 5.1 also shows higher mean values for higher-scored events. These results demonstrate numerically the effect of morphologic changes of IEDs on expert scoring.

Figure 5.7: Waveforms with score: 3.3-5 sorted from low to high average score. Each waveform lasts 0.6 s at the same amplitude scale.

Table 5.1: Mean values for morphological features of all events, and events scored 4, and above it.

Group of events	Peak voltage (μV)	Rising edge voltage (μV)	Falling edge voltage (μV)	Peak to peak voltage(μV)
All events	49.185	81.810	94.973	107.68
Events scored 4 and higher	70.171	129.733	142.88	165.933
p-value	0.001	4.95e-07	3.13e-07	2.17e-08

5.3.3 Visualization by t-SNE

In this section, the waveforms are visualized by t-SNE method, using all the features extracted for the events.

Figure 5.8: Box plots of morphological characteristics of IEDs for all events, and events scored ≥ 4 .

T-SNE plots for all events using all extracted features are shown in Fig. 5.9 and Fig. 5.10. Each waveform is color coded based on its assigned average score given by either all or top 7 raters, respectively. As shown in Fig. 5.9 and Fig. 5.10, by means of t-SNE we are able to visualize the feature array into a 2D domain.

We observe that the events with different scores are mapped to different locations in two dimensions. It appears that “spikiness” is found on the y-axis (D2) and the presence or absence of a slow wave is present on the x-axis (D1). Furthermore, this plot shows that it is possible to visually discriminate between the events scored on a 5-point scale, by employing the set of extracted features. The t-SNE using top scorers seems to provide slightly better separation between events.

Figure 5.9: t-SNE plot using all features, labelled based on all 18 scorers.

5.3.4 Expert scoring results

19 EEGers used the EEGnet system to label a total of 1,115 single-channel segments of the EEG as containing IEDs, in the first phase. The average number of IEDs labeled by the EEGers was 58.7 (range, 16–195). A clustering algorithm was applied to create a master list of 235 segments to remove redundancy and to include only events which were marked by at least two EEGers, as described above. This phase 1 scoring task was completed by each EEGer within 1 to 2 hours. In the phase 2 scoring task, the 19 EEGers were asked to categorize the events in the master list of EEG segments on the 5-point Likert scale described above and 18 EEGers completed this task, making a total of 4,320 classification decisions. In phase 3 scoring, each EEGer was presented with a unique list of between 30 and 104 events to be rescored, if there was a difference between their phase 1 and phase 2 scoring opinion. The scoring results considered for analysis were the phase 2 results corrected for consistency with the phase 3 results.

The IRA for all EEGers was 0.31 based on Gwet’s AC2, indicating low to mod-

Figure 5.10: t-SNE plot using all features, labelled based on top 7 scorers.

erate agreement, as found in previous studies of IED detection [45, 48, 79, 134, 137]. In addition, there was a high degree of marginal inhomogeneity (mean pair-wise Stuart-Maxwell statistic: 76.7, range 1.7 – 176.6). For example, the average EEGer marked 19% of events with a score of 5 but this varied between 7-51% among EEGers.

We calculated the IRA of all possible groups of EEGers ranging in size between 2-17 expert scorers (262,124 groups). The AC2 IRA for the group with the highest agreement, as a function of group size, as well as the AC2 IRA of the remaining EEGers, are shown in Fig. 5.12. Fig. 5.11 demonstrates the IRA among randomly selected subsets of EEGers with different size, for which we do not observe any pattern in IRA of selected and remaining scorers.

We observe that there are groups of EEGers of various sizes who agree most among themselves, while the rest of the EEGers have a low degree of agreement. We observe From Fig. 5.12 that as the group size increases, the IRA of the group with the maximum agreement decreases. We call the subgroup of EEGers of size (n) having the highest AC2 IRA the “top scorers”.

Figure 5.11: AC2 agreement coefficient among the groups of randomly selected scorers of size n .

Figure 5.12: AC2 agreement coefficient among the groups of top scorers of size n .

5.3.5 Features versus expert scoring

We performed a correlation analysis on all the features which allowed us to compare feature categories based on how much they correlate with expert scores. This is important since a large number of features are considered in this analysis, and by applying elastic net feature selection we can remove most of the redundant and highly correlated features. Therefore, while we considered feature selection output for machine learning analysis, we also performed correlation analysis independent from feature selection, with the purpose of having an overall comparison between all considered features.

Fig. 5.13 demonstrate the Spearman correlation between each group of features with $\bar{S}$ and σ_s . The correlation values correspond to the maximum correlation for a feature in each category to provide an overall comparison between various feature types. We observe that wavelet coefficients as a category have the highest correlation with both average and standard deviation of scores.

Figure 5.13: Feature categories and their maximum correlation with the average and standard deviation of the expert scores.

The top features of each category are listed in Table 5.2 and Table 5.3 for $\bar{S}$ and σ_s , respectively. Most of them correlate positively with the mean scores, includ-

ing magnitude and standard deviation of wavelet coefficients and NLEO, energy, PSD, morphological features, and line length. An exception is waveform duration, which is inversely correlated with the average expert score. By contrast, most of the top features are inversely correlated with the standard deviation of expert scores, as shown in Table 5.3. The electrode location features will be discussed in Section 5.3.6.

Table 5.2: Highest correlation between the features from each category with the mean expert scores. ψ represents wavelet basis function, and ϕ_k is the NLEO with resolution parameter (k). Subscripts ' t ' and ' f ' indicate the operation on PSD values along time or frequency axis, respectively.

Wavelet		NLEO		Morphological		Energy		Spectrogram	
$\psi = dmey$	0.669	$\sigma(\phi_{k=6})$	0.559	Peak to	0.593	12-20 Hz	0.595	$\sigma(P\bar{S}D_f)$	0.373
$ min(d3) $				peak voltage					
$\psi = sym5$	0.667	$max(\phi_{k=6})$	0.542	Rising	0.587	12-30 Hz	0.579	$\sigma(P\bar{S}D_t)$	0.333
$max(d3)$				voltage					
$\psi = bior5.5$	0.666	$max(\phi_{k=11})$	0.539	Rising	0.572	10-30 Hz	0.569	$P\bar{S}D$	0.317
$max(d3)$				slope					
$\psi = rbio5.5$	0.665	$max(\phi_{k=16})$	0.537	Peak voltage	0.570	10-42 Hz	0.556	$Max(PSD)$	0.306
$max(d3)$									
$\psi = db8,$	0.663	$\sigma(\phi_{k=1})$	0.533	Falling	0.568	8-34 Hz	0.547		
$max(d3)$				slope					
$\psi = coif4,$	0.660	$max(\phi_{k=21})$	0.512	Falling	0.566	6-40 Hz	0.541		
$max(d3)$				voltage					
$\psi = rbio3.7$	0.659	$\sigma(\phi_{k=16})$	0.501	Duration	-0.204	4-38 Hz	0.174		
$max(d3)$									

Table 5.2 shows that the wavelet basis functions most correlated with $\bar{S}$ include *dmey*, *sym5*, *bior5.5*, *rbio5.5*, *db8*, *coif4*, *rbio3.7*; and wavelet coefficients of *d3* decomposition scale have the highest correlation with $\bar{S}$. In addition to these features, scales of *d4*, *d2*, *a2*, *d1*, *a1*, and *d5* come next, respectively. From Table 5.3, it is observed that *sym4*, *sym5*, *rbio5.5*, *rbio4.4*, *coif2*, *rbio2.4*, *db6* have the highest correlation with σ_s .

Table 5.3: Highest correlation between the features from each category with standard deviation of expert scores. ψ represents wavelet basis function, and ϕ_k is the NLEO with resolution parameter (k). Subscripts ' t ' and ' f ' indicate the operation on PSD values along time or frequency axis, respectively.

Wavelet		NLEO		Morphological		Energy		Spectrogram	
$\psi = sym4$	-0.513	$\min(\phi_{k=1})$	0.382	Rising	-0.365	6-50 Hz	-0.351	$\sigma(P\bar{S}D_f)$	-0.218
$ min(d1) $				slope					
$\psi = sym5$	-0.502	$\sigma(\phi_{k=1})$	-0.345	Rising	-0.364	4-50 Hz	-0.344	$P\bar{S}D$	-0.217
$max(d1)$				voltage					
$\psi = rbio5.5$	-0.497	$max(\phi_{k=1})$	-0.267	Peak to	-0.295	6-32 Hz	-0.344	$\sigma(P\bar{S}D_t)$	-0.212
$max(d1)$				peak voltage					
$\psi = rbio4.4$	-0.497	$\min(\phi_{k=16})$	0.250	Falling	-0.280	8-46 Hz	-0.315		
$max(d1)$				voltage					
$\psi = coif2$	-0.492	$max(\phi_{k=31})$	-0.244	Peak	-0.247	8-32 Hz	-0.312		
$max(d1)$				voltage					
$\psi = rbio2.4$	-0.491	$max(\phi_{k=26})$	0.241	Duration	0.207	4-16 Hz	-0.308		
$max(d1)$									
$\psi = db6$	0.491	$max(\phi_{k=21})$	0.237	Falling	-0.162	2-50 Hz	-0.265		
$min(d1)$				slope					

In addition, wavelet coefficient for decomposition scale of $d1$ have the highest correlation with σ_s , followed by $d4$, $d2$, $a4$, $d5$, $a1$, $a2$ and $a3$, respectively.

5.3.6 Electrode position on scalp

Spearman correlation of the electrode location coordinate features based on a two-dimensional projection of the scalp electrode positions demonstrated a weak correlation between expert scores and several electrode location features. The correlation coefficients between the scores and electrode location features are listed in Table 5.4. Experts' scores are greater anteriorly than posteriorly ($\rho = 0.32$), indicating that the EEGers tended to label

anterior events as more epileptiform than posterior events. Higher score events tend to be more in non-temporal regions ($\rho = -0.35$) and in the central versus lateral regions ($\rho = -0.42$). There is not a significant relation between electrode coordinates and the standard deviation of the scores (σ_s).

Table 5.4: Correlation coefficient between electrode coordinates, and mean and standard deviation of expert scores.

Electrode location comparison	Correlation with mean expert scores	Correlation with standard deviation of expert scores
Anterior vs. posterior	0.32	0.07
Lateralization (left vs. right side)	-0.31	-0.09
Temporal vs. non-temporal	-0.35	-0.10
Lateralized vs. central	-0.42	-0.16

5.3.7 Feature selection

In order to visualize the relation between features and scores, scatter plots of some selected features after feature selection versus the average score of all scorers $\bar{S}_{all}$ are shown in Fig. 5.14 and Fig. 5.15. Due to space constraints, we only show features from each category of selected features. The fitted regression lines, as well as Pearson correlation coefficient which quantifies linear relationship between features and scores, and Spearman coefficient which shows nonlinear relationship, are shown on the plots as well. It is observed that in these plots, the feature value will increase as the score increases, indicating certain features do have significant impact on expert scores. In these plots, the feature value increases as the score increases, indicating certain features do have significant impact on expert scores. Features selected using elastic net regression which

Figure 5.14: Scatter plots of some chosen features after feature selection versus expert scores. Pearson correlation values between features and scores are shown on the plots.

Figure 5.15: Scatter plots of some chosen features after feature selection versus expert scores. Spearman correlation values between features and scores are shown on the plots.

were highly predictive for $\bar{S}$ and σ_s , are presented in Fig. 5.16 and Fig. 5.17, respectively.

We selected the parameter $\alpha = 0.8$ for the feature selection by cross-validation. In the

figures, the features are rank-ordered based on their absolute coefficient in the elastic net feature selection. We applied 10-fold cross validation during feature selection; and all features presented in Fig. 5.16 and Fig. 5.17 were selected in all 10 cross validation folds. It is important to note that the selected features are not necessarily the only beneficial features for predicting expert scores. During feature selection, highly correlated features are eliminated; and by running feature selection multiple times, or by making slight changes to the signals such as shifting, the resulting set of features may be slightly different. From Fig. 5.16 we observe that mainly wavelet features are selected to predict

Figure 5.16: Features selected in feature selection step, which are more significant in predicting the average scores. ψ represents wavelet basis function, ϕ_k is the NLEO with resolution parameter (k).

mean expert scores. Morphological features such as rising and falling voltages and duration, electrode location (lateralization and anteriority), NLEO features, and peak voltage are also selected though they receive less weight. The features with negative coefficients

Figure 5.17: Features selected in feature selection step which are more significant in predicting standard deviation of the scores. ψ represents wavelet basis function.

correlate inversely with average scores, for which lower feature value corresponds to a higher score.

We also observe in Fig. 5.17 that most features selected as predictors of the standard deviation of scores are a different set of wavelet features, followed by energy in 10-30 Hz, electrode location (lateralization) and morphological features including rising voltage and slope, and duration. For these features also, negative coefficients indicate negative correlation, and higher values of the features with negative coefficients correspond predict lower standard deviations in expert scores.

In addition, we observe from Fig. 5.13 and correlation values listed in Table 5.2 and Table 5.3, that wavelet features have the highest correlation with both mean and standard deviation of expert scores. Moreover, among the selected features mentioned in Fig. 5.16 and Fig. 5.17, wavelet features have the largest coefficients. Hence, wavelet features are shown to be more beneficial in characterizing expert scores in IED interpretation. To make these results more tangible, the morphology of the wavelet basis functions

which are shown to be more useful are provided in Fig. 5.18. Nevertheless, other features such as morphological features are also seen to be highly correlated with the scores, thus may be helpful in IED interpretation.

Figure 5.18: The wavelet basis functions listed in Table 5.2 and Table 5.3 which have the highest correlation with the mean and standard deviation of the expert scores.

It could be expected that wavelet features perform better than other features in this study. Analysis of electrical signals usually calls for some type of time-frequency analysis. Wavelet analysis is an efficient method for doing this and wavelet signal processing techniques have recently led to significant gains in critical signal processing tasks. In this method, data is split into different frequency components, and each component is studied with a resolution which matches its scale. If the analysis approach is well-matched to the task, wavelets provide an extremely sparse and efficient representation for several types of signals appearing often in real world applications, but are not well matched by the Fourier basis, which is ideally meant for periodic signals. In particular, they provide an optimal representation for many signals containing singularities (jumps

and spikes) [167], of which IEDs are a classic example.

5.3.8 Multivariate analysis of expert agreement

We fitted support vector regression (SVR) models to predict the expert scores using the features selected above and evaluated the performance by 10-fold cross-validation.

We fitted SVR models to the dataset for the scenario in which all scorers are included, and for top scorers of various group sizes (17 groups in total). The target variable to be predicted by the model is the average score given to each event by each of the 17 expert groups (the group of all experts and the 16 “top scorer” groups). This was done separately for both $\bar{S}$ and σ_s . The resulting RMSE values and the bars indicating $+/-1$ standard deviation from the 10-fold cross-validated SVR models are shown in Fig. 5.19. We see that, as the number of top scorers increases, error in predicting expert agreement decreases. Performance is best when using the average score of all available 18 EEGers. The performance appears to level off when the top scorer group size increases to 12. We would not expect the SVR performance to improve much after 12 top scorers have been added, because the mean expert opinion in these large top scorer groups above size 12 does not change much as more EEGers are added [150]. Nevertheless, because the group of all scorers produced the best overall performance, we choose “all experts” to be the optimum set of experts to be considered in the further analysis presented below. We considered three different sets of feature inputs to train the models: (1) raw EEG waveforms, (2) all extracted features, (3) and selected features using elastic net regression. Furthermore, we compared the models using non-normalized and normalized data types. RMSE values for these SVR models are listed in Table 5.5. In addition to the waveforms centered on the peaks, the errors for waveforms shifted by a random number of samples ranging from 1 to 10 are included in Table 5.5. Table 5.5 also shows the errors for the SVR models trained on the feature set with DT-CWT. We observe that although shifting the EEG signal changes prediction errors, the effects on the prediction error are small when using the set of extracted wavelet statistics. This is probably because the wavelet transform coefficients are not used directly; instead, statistics of the coefficients within a small neighborhood around the peak are considered. Therefore, the features utilized in this study have some shift invariance built in. In addition, no significant change

Figure 5.19: Root mean square errors resulted from SVR models fitted to the dataset, considering groups of top scorers of different sizes ($n = 2$ to 18); $n = 18$ indicates the group of all experts.

was observed in the feature values and set of selected features by shifting the signals. Moreover, the error values when considering the DT-CWT as an alternative to DWT, are higher than the feature set including the DWT. This may be because the four wavelet filters applied in DT-CWT are not appropriate for IED analysis. Consequently, we consider the DWT features to be a better choice of wavelet transform for our analysis. The smallest SVR error is achieved by utilizing the set of elastic net selected features on the data with local standardization. For this model, the predicted scores have a high correlation ($\rho = 0.87$) with the mean expert scores, as shown in Fig. 5.20, with the RMSE and NRMSE performance values for the predictions 0.46 and 0.50, respectively. After fitting an SVR model to predict the standard deviation of scores using the selected features, the correlation coefficient is 0.62, whereas the predictions have RMSE and NRMSE of 0.21 and 0.79, respectively. Fig. 5.21 shows a Receiver operating characteristic (ROC) curve for this model. In this figure, the ‘sensitivity’ (true positive rate) is plotted versus ‘1-specificity’ (false positive rate). ‘Precision’ versus ‘sensitivity (recall)’ is demonstrated

Table 5.5: RMSE values of SVR models based on average expert score for all EEGers for normalized and non-normalized data types, as well as various sets of inputs to the algorithm. The errors are calculated both for predictions made by peak-centered and shifted signals.

Data Type →	Raw EEG waveforms			All features			Features selected with elastic net regression		
	DWT Centered	DWT Shifted	DT-DWT Centered	DWT Centered	DWT Shifted	DT-DWT Centered	DWT Centered	DWT Shifted	DT-DWT Centered
No normalization	0.80	0.85	0.87	0.58	0.59	0.64	0.49	0.49	0.57
Local standardization	0.85	0.88	0.86	0.61	0.61	0.63	0.47	0.49	0.56
Global standardization	0.81	0.83	0.88	0.67	0.68	0.72	0.61	0.60	0.66
Median background RMS normalization	0.80	0.83	0.88	0.67	0.68	0.70	0.63	0.62	0.66

in the precision-recall curve in Fig. 5.22. To create the ROC and precision-recall curves, we first set the ground truth. The events scored 4 or 5 by at least half of the experts (at least 9 experts) were considered as IED, while the other events were labeled as non-IED. Next, we applied a threshold θ ranging from 1.5 to 5 to the predicted scores of test waveforms. We classified the events with scores of higher than θ as IED, and we classified events with lower scores as non-IED. The area under curve (AUC) equals to 0.94 for the ROC curve, and 0.92 for the precision-recall curve.

5.4 Conclusions

Identification of IEDs in EEG by experts is a cornerstone of the medical diagnosis of epilepsy, yet the features which mark an EEG waveform as “epileptiform” are difficult to specify with precision. The International Federation of Clinical Neurophysiology defines IEDs [28] as "distinctive waves or complexes, distinguished from background activity, and resembling those recorded in a proportion of human subjects suffering from epileptic disorders...." This highlights the fact that expert recognition of IEDs is based on experience rather than rigid definitions and helps to explain imperfect agreement between

Figure 5.20: Predicted scores produced by SVR trained using elastic net selected features versus original expert scores, using EEG signals normalized by local standardization

experts scoring the same EEG. In the present study we evaluated numerous EEG features to determine which ones best predict whether a waveform will be judged by experts to be an IED, and which features best predict differences of opinion. We found that expert opinion regarding IEDs is predictable when the number of experts is large.

We also found that the most effective features to consider in models that predict higher expert scores were wavelet features (using certain wavelet basis functions, as listed in Fig. 5.16). Certain features (as listed in Fig. 5.17) were also associated with a high level of variance in scores, indicating disagreement among EEGers, although it was not possible to predict the variance among scorers as well as the mean score.

As expected, as the number of experts interpreting EEG increases, IRA decreases, as would be expected by chance. Although the overall IRA of the EEGers was

Figure 5.21: ROC curve for classification based on the SVR output.

only mild to moderate, there was a subset of EEGers with good IRA. Groups of top scorers were found which have substantially higher IRA, which raised the question of whether using just a subset of top scorers rather than all scorers to train the SVR models might improve model performance. However, based on prediction error for subsets of top scorers with different sizes, we found that the consensus opinion of all available experts, rather than a subset with good IRA, resulted in the most accurate predictions. A “wisdom of the crowd” (WOC) effect seems a likely explanation of this observation: The aggregated answers of a large group are frequently better than the answer given by any individuals within the group [181, 182]. WOC effects are observed when individual judgments can be thought of as “noisy” but are correlated with an underlying, true signal (as opposed to having no objective basis), such that averaging tends to cancel the noise. Nevertheless, the performance of our models levels off at a top scorer group size of around 12. This is consistent with a detailed analysis of our expert opinion published elsewhere which showed that the average opinion of our experts ceases to change significantly when the

Figure 5.22: Precision-recall curve for classification based on the SVR output.

top scorer group gets to the size of 10-12 [150].

The FDA currently only requires three EEGers to annotate the location of IEDs in a testing dataset for approval of an IED detection algorithm. Most studies of IED detection algorithms in the past have expert opinion from five or fewer experts [3] and previous work by Wilson *et al.* suggested that probably only three EEGers were needed to annotate EEG datasets [134]. The present study suggests that using more than three EEGers will probably produce better results.

Since average amplitudes of EEG signal and artifacts vary between individuals, we hypothesized that it might be helpful to normalize EEG signals before feature extraction. We observed better prediction is achieved if normalization is performed by locally standardized EEG by z-scores. Globally normalizing each channel of EEG data did not improve the prediction error. This is a useful finding since it may inform best practices regarding preprocessing steps for future IED detection algorithms.

We also investigated the relationship between expert scores and the location of

EEG events on the scalp. The data suggests that experts tend to give higher (more epileptiform) scores to events which are: (1) anterior rather than posterior, (2) more in the non-temporal region rather than temporal regions, and (3) more central compared to lateralized. This result may not reflect a finding about IEDs in general since it may only be due to selection bias, since only part of our EEG dataset was selected randomly. But the finding of more non-epileptiform paroxysmal activity in the temporal region agrees with reports that benign epileptiform-appearing events are frequently found in the temporal regions [183].

This study is subject to several important limitations. First, the number of EEG events considered is probably quite a bit lower than the total number of EEG events necessary to completely encompass the variability found in epileptiform and benign epileptiform-appearing signals in all individuals. Future definitive studies will need to include larger EEG datasets and more subjects. Second, the IRA of our expert opinion was only low to moderate, calling into question the quality of the expert opinion considered to train the machine learning models of this study. IRA between experts might have been higher if the EEG segments from each subject had been longer than 30 seconds. It is known that IRA is higher if experts are interpreting 20-minute routine EEG recordings than shorter EEG recordings [58]. It should be noted that the IRA was not uniformly low among our EEGers, since there was a large subset of EEGers with moderate to high IRA. This illustrates the importance of carefully selecting EEGers for IED annotation research projects. Third, the high degree of marginal inhomogeneity in the expert opinion of this study is a weakness and may have affected results. Hopefully, this can be decreased in future studies by providing more detailed instructions to EEGers. Fourth, although this study suggests certain features that may be beneficial in constructing IED detection algorithms, creating and testing an IED detection algorithm was beyond the scope of this study. In particular, the present study does not address how well our features might discriminate IEDs from background waveforms when large amounts of EEG are examined in a typical IED detection system, a situation in which non-IED background waveforms vastly outnumber IEDs.

In conclusion, the results of this study show that expert opinion regarding agreement or disagreement regarding the interpretation of IEDs is predictable. In addition, our

analysis has identified preprocessing steps and particular features which are effective for training machine learning algorithms to predict expert agreement regarding IEDs, including local normalization of waveforms, and particular types of wavelet functions, among other features. Our analysis reveals that aggregated scores from a large number of experts are substantially more predictable than scores from small numbers of individuals. This finding suggests that despite studies finding that inter-rater agreement among experts is imperfect (a) there is an underlying objective and predictable reality that can be captured by machine learning algorithms, and (b) the common practice of seeking consensus by group review on difficult EEG cases may, in fact, be helpful in getting closer to the truth. It is our goal to create a large freely-available standardized testing and training database for the improvement of IED detection algorithms and work to improve IED detection algorithms themselves. In the future, we hope to collect expert opinion on longer EEG recordings and to develop a complete IED detection system.

Chapter 6

Computers vs Human

Electroencephalographers for Recognizing Epileptiform Discharges

In this chapter, we devise and implement methods for evaluation of a IED detection algorithm. Moreover, we investigate an existing IED detection software and analyze the outcome.

6.1 Introduction

Automated IED detection systems have employed multiple methods as described in Chapter 2 and have been under development for more than 50 years [3, 29]. Despite past and ongoing efforts, there are still no universally accepted automated IED-detection systems available.

In a recent study by Scheuer *et al.* [137], performance of the commercial software Persyst 13 (P13) [184] for IED detection on a dataset consisting of 40 epileptic patients was compared with IED detection by three EEG technicians. In that study, 100

consecutive EEG recordings containing recorded electrographic seizures from 100 different individuals were identified. From that initial set of 100 EEGs, a subset of recordings which was reported to contain IEDs at a rate greater than one per hour was generated. 35 records of this subset were randomly chosen for detailed IED annotation. In addition, another subset of recordings which was reported to be IED-free was generated from the 100 initial recordings; and 5 of these records were randomly selected as IED-free or low IED-rate segments; they were combined with the records which contain IEDs. The dataset presented to readers consisted of 40 records, with 35 records containing IEDs and 5 IED-free records. Three senior EEG technologists were chosen to mark the test EEG data set for IEDs (spikes and sharp waves). The technologists were instructed to begin assessing each EEG at its beginning and to “carefully mark every epileptiform spike or sharp wave encountered, per the standard definition and their clinical experience”, up to 100 IEDs. The P13 detector was run on these EEGs, which produced a list of IED detections, each with an associated perception value (PV), which indicated how confident the detector was that the detection was an IED. At a perception threshold value (PTV) of 0.1, it is reported that the P13 algorithm passed a “statistical Turing test”, meaning that for IED detection, human-algorithm inter-rater agreement (IRA) was comparable to IRA between the technicians. However, there are limitations to this study, as explained by Westover *et al.* [185]. Most importantly, validation of these results on a larger set of EEGs and comparison with physician-level EEG specialists is needed. In the Scheuer *et al.* study [137], the three human annotators were EEG technologists, and it is unclear how well the P13 algorithm would perform versus clinical neurophysiologists, who are the interpreters of EEG in clinical practice. In this paper, we compare the performance of the P13 algorithm to that of 19 academic clinical neurophysiologists. Moreover, the methodology used to compare human vs algorithm performance in the Scheuer *et al.* study [137] is problematic, as we will show herein.

In comparison to development of traditional diagnostic testing for epilepsy, such as human interpretation of EEG and scans such as MRI, the development of FDA-approved tools for automated detection of neurophysiologic signals to aid in clinical neurophysiology practice is a recent phenomenon [137, 186, 187]. Little research has been done to study methods for acquiring expert opinion to test clinical neurophysiology automated

event detection systems. The FDA currently requires three experts to review each recording in standardized testing databases for automated neurophysiology event detector systems, but these experts are not required to be physicians, and the level of training of the experts is not precisely specified. The choice of three experts seems arbitrary and the lack of precise specification for the level of training of the experts seems problematic, and likely flows from the paucity of objective evidence to support the need for qualifications for the experts. In this study, we explore the performance of the P13 detector versus a large number of neurologists and propose an approach for objective measurement of the size of the expert group and the qualifications of an expert group to test IED detectors.

6.1.1 Persyst IED detection software

In the paper published by Persyst Development Corporation [137], the P13 version of Persyst software is claimed to be non-inferior to humans in the task of detecting spikes in interictal EEG of patients with epilepsy. In their paper, 3 human annotators have been considered, and none of these 3 annotators are neurophysiologists. It is unclear how well these results might generalize to larger, more comprehensive sets of EEGs. Moreover, there was substantial disagreement among human experts marking the dataset.

The overview of the algorithm applied by the P13 is as follows.

The IED morphology was represented by a six-half-wave model, with the middle two half-waves describing the traditional negative/positive deflection of the IED, and the last two half-waves describing the slow wave, if it exists, and the first two half waves describing the activity immediately preceding the IED. These half-waves were also used to discriminate IEDs from rhythmic activity. Each half-wave was described by its amplitude, duration and curvature. The vertex between each pair of half-waves was described by its angle and tilt from vertical. Approximately 20 feedforward neural network rules (NNrule) were used. Each NN-rule comprised an ensemble of up to 40 individual NNs, each one having 1 to 3 hidden layers. Each NNrule was designed to encapsulate an expert-derived concept, e.g. “if the amplitude and sharpness of the event exceeds that of the background activity, then the likelihood that the event is a spike is increased.” They were then hierarchically nested to propagate uncertainty and to allow the poor candidates

to be pruned before performing computationally expensive operations. 3 thresholds were applied in the hierarchy to speed processing by removing events that are clearly not IEDs. A perception value is assigned to un-pruned events, visible to the user. This value varies between 0 and 1. A perception value near 0.5 is assigned to uncertain events.

Three senior EEG technologists were chosen to mark the test EEG data set for spikes and sharp waves. They were instructed to assess each EEG at its beginning and mark IEDs per the standard definition and their clinical experience, up to 100 spike marks. If the annotations from two readers, human or algorithm, coincided within 200 ms, they were considered a match. Nonmatching detections were considered as false positives. Forty EEGs were marked by the 3 readers. The mean length of the original recordings was 23.2 h (6.4–24.2 h.). A total of 253 h of EEG were marked by all three readers.

The human readers had only a fair degree of agreement with one another, with an average of around 45% sensitivity for one another's IED marks. Using an accelerated bootstrap test as a test of possible non-inferiority (NI), the Persyst 13 IED detection performance was claimed to be statistically non-inferior to the skilled human readers, while the Persyst 11 Reveal and Persyst 12 were reported to be inferior. Fig. 6.1 shows the ROC curves for P13 for IEDs marked by minimum one, two, or three readers.

Figure 6.1: False positive rate versus sensitivity for the P13 software for three levels of experts agreement [137].

6.2 Materials and Methods

6.2.1 Dataset of EEG recordings and expert opinion

200 thirty-second segments from EEG recordings from 200 different patients at the Medical University of South Carolina were chosen by two board certified clinical neurophysiologists (who were not used as expert scorers). Selection of segments was carried out as follows:

- Fifty segments were from within EEGs that were interpreted as containing IEDs in the clinical report.
- Fifty segments were from within EEGs that were interpreted as normal in the clinical report
- Fifty segments were chosen to contain IEDs from epilepsy patients with difficult-to-interpret IEDs.
- Fifty segments were chosen from patients with benign epileptiform-appearing paroxysmal activity which an inexperienced reviewer might misinterpret as being epileptiform.

All 200 segments were randomly ordered and concatenated into one long 100-minute recording. Experts annotated EEG recordings using EEGnet, a software system hosted at the School of Computing of Clemson University. This software allows users to view and annotate EEG datasets on a web browser without the need for locally installed software [7]. EEGnet is designed to provide a user interface similar to commercially available software applied in clinical EEG interpretation. Ten seconds of EEG data are displayed at one time. Furthermore, channel labels along with one electrocardiogram channel are displayed. Channel gain can be adjusted for all channels or for individual channels, and users can scroll through the signal one or ten seconds at a time. Users can adjust digital high pass, low pass, and notch filters. The sweep was fixed at 10 seconds. Several conventional EEG montages including hatband bipolar, transverse bipolar, anterior-posterior bipolar, Cz reference, ipsilateral ear reference, and common average reference are available [7].

As per the scoring protocol for the study [188], the experts annotated IEDs by labeling a short epoch of time on a single channel, on one of a range of montages, which the expert thought provided the best representation of the IED. Because analysis of IRA requires matching of IED annotations between experts, clustering was performed to determine which of the single channel time epoch annotations by a given expert corresponded to annotations by other experts. A clustering algorithm was applied to eliminate redundancy by merging labeled EEG segments that strongly overlap in time [7]. If two labeled EEG segments overlapped in time by at least half of the duration of the shortest segment, they were placed into the same cluster. The representative EEG segment for a cluster (a “cluster center”) was defined as the EEG segment which had the maximum sum of correlations with all segments of its cluster. For each cluster center in our dataset, the start time was defined to be the first start time found among the start times of all the annotations in that cluster. Similarly, the end time was defined as the last end time among all the annotations in that cluster. If the peak of a detection by P13 fell between a cluster’s start and end time, it was considered to be a detection which matched that cluster of human expert annotations.

Annotation by 35 neurologists resulted in 4297 annotations and subsequent clustering produced 573 cluster centers. A sample of an IED within a 10 s page of EEG rendered in the EEGnet system is presented in 6.2. Among the 35 neurologists, 19 were academic clinical neurophysiologists. A subset of 23 of 35 were board certified in clinical neurophysiology by either the American Board of Clinical Neurophysiology (ABCN) and/or a relevant American Board of Psychiatry and Neurology (ABPN) subspecialty exam (ABPN Clinical Neurophysiology and/or ABPN Epilepsy). A total of 25 of 35 experts had completed at least one year of clinical neurophysiology fellowship training. The annotations of these experts were compared with P13 detections. For the majority part of the analysis, the 19 clinical neurophysiologists are considered for comparisons, unless otherwise stated. The P13 algorithm provides a PV for each of its IED detections, within the range of 0.1 (most sensitive) and 1 (least sensitive), which allows the user to adjust the sensitivity of the detector, since the user can adjust the PTV so that only detections above a specified PV are shown to the user as valid detections. The performance of the P13 detector can be measured at different settings of the PTV. For this study, all data

Figure 6.2: A sample EEG segment considered in the current dataset, displayed and annotated by EEGnet software.

processing and statistical analyses were performed offline using a commercial software package (MATLAB R2016b).

6.2.2 IRA using Gwet's AC2

Gwet's agreement coefficients, AC1 and AC2, [161, 162] are statistical measures of IRA for two raters and for more than two raters, respectively. These measures behave better than Cohen's kappa and Fleiss' kappa in cases of low marginal homogeneity [189]. A comparison of these methods was also provided in Chapter 5. In this study, we computed the linearly weighted Gwet's AC2 agreement between the P13 scores and human expert scores as the method for calculating IRA .

6.2.3 Receiver operating characteristics (ROC) based on expert consensus

We computed a sensitivity (Se) and false positive rate (FPR) for each expert relative to consensus among at least n experts, for various values of n . For P13, because different values of sensitivity (Se) and false positive rate (FPR) are obtained for each setting of PTV, we calculated a series of paired (Se, FPR) values, i.e. an ROC curve, by varying

PTV from 0 to 1. When calculating (Se, FPR) for an individual expert, the scoring data from that expert was left out of the calculation. For instance, when calculating the performance of an expert based on the consensus of at least 3 experts, all waveforms annotated by at least 3 experts (not counting that expert) are considered as true positive (TP) annotations, and the rest as false positive (FP) annotations. ROC analysis was performed for four scenarios with differing fractions of expert consensus among 19 academic neurophysiologists (consensus among $n = 5, 8, 12$, and 15 experts). We calculated P13's ROC performance similarly, except that all 19 human experts were included in the calculation.

6.2.4 Mean pairwise ROC analysis

We carried out mean pairwise comparisons between P13 and each expert with respect to all other experts, and generated an ROC plot using the pairwise performance metrics. We calculated pairwise performances such that for every pair, the annotations of each expert are considered as the ground truth and the mean pairwise sensitivity and false positive rate ($Se, 1-Sp = FPR$) with all remaining experts is computed. For P13, we calculated the mean average performance with all experts.

6.2.5 Bootstrap statistical analysis of mean pairwise ROC performance

Since the Sheuer *et al.*'s paper [137] reported that the P13 detector was non-inferior in performance to a group of three human experts at a PTS of 0.1, it is interesting to measure non-inferiority in performance for the P13 using our EEG testing dataset and subsets of our neurologist expert opinion. This would allow us to examine the effect of expert group size and expert group composition on performance testing of the P13 detector. This might help us determine the probability of the P13 detector being found to be non-inferior in performance to a group of experts, based on the size and composition of the expert group.

We applied a non-inferiority statistical test similar to the test used by Sheuer *et al.* [137] to the present dataset. In the study by Scheuer *et al.*, a test of possible non-inferiority of either P11 [75], P12, or P13 algorithms to the human readers, was performed. This test was applied by a bias-corrected and accelerated bootstrap analysis

(BCa, $N = 3000$) of the pairwise differences of sensitivity and FPR for each EEG record. The term non-inferiority [190] is used to denote the absence of a statistically significant difference in performance between the human readers and the automated detector.

Considering the Sheuer *et al.* approach [137], we performed a BCa in the present study to find the 0.95 confidence interval for the average of each pairwise performance measurement difference, resulting in one performance measurement difference for each pair. By considering all two-expert combinations among a pool of $n = 19$ experts, and calculating the pairwise performance (for either pairwise set of (Se, FPR) values) between each pair of experts, this produced $x = 171$ expert pairs among the entire pool of experts, and $2x = 342$ performance difference measurements. The maximum allowed performance difference measurement for another additional expert (whether they are a human expert or automated detection system), if added to this group of n experts, if that additional expert is to be considered as non-inferior in performance when compared to the existing group of experts, based on bootstrapping, was found based on all bootstrapped performance difference measurements. The acceptable decrease in sensitivity for the pairwise difference between the group of n experts and the additional expert is denoted by δ_{sens} . The acceptable increase in false positive rate between the group of n experts and the additional expert is defined as δ_{FPR} .

To determine whether P13 was non-inferior in performance to human experts, we tested the detector using this non-inferiority statistical approach by treating P13 as an additional expert. We computed the 0.95 confidence intervals for all P13 versus human expert pairs. If all the pairwise performance difference measurements (by finding the 0.95 confidence interval for mean performance difference measurement) are within the maximum allowed difference, the detector is considered to be non-inferior to humans.

We calculated the Se and FPR for each record resulting in 200 pairs for each round of bootstrap analysis. In the calculation of sensitivity, if for any PTS total number of IEDs in a record is zero, $Se = 1$ if

the number of true positives + false negatives + false positives = 0, otherwise, $Se = 0$.

The ability to measure the performance of an automated detector, which is trained to mimic human expert detection performance, is assumed to increase as the number of test datasets is increased. If the detector also provides a PV with each detection (as the

P13 detector does), it is possible to measure the optimal range for the PTS. We have attempted to model this using the P13 detector, our dataset of short EEG samples, and our expert opinion. Given an EEG dataset constructed to test the performance of an automated detection system, consisting of nE number of EEG samples drawn from nE different human patients, EEG subset datasets can be created by random sampling with replacement from iE of the total nE number of EEG samples.

We tested the IED detection performance of P13 versus human expert performance across the range of potential PV threshold settings. The non-inferiority statistical test described above declares the P13 detector non-inferior to human experts for any given PTS if the difference for both pairwise performance metrics for P13-human pairs is within the maximum allowed difference defined by bootstrapping.

Therefore, in the present study, we assessed P13 detector using this non-inferiority test for a detector set, at a particular PTV (St) versus a group of experts G . This non-inferiority test is expressed as an operation with a binary output, $NI[G, St]$. We performed this analysis in two directions explained in the remainder of this section.

6.2.5.1 Investigating the effect of PTS and number of EEG subjects

This can be modeled using a two-dimensional matrix M constructed from binary non-inferiority values $NI[G, St]$ using as input an appropriately-sampled range of y-axis St values and various sizes of the sampled EEG testing dataset varying from 1 to nE . The calculation of each value in M can be repeated z times for smoothing, with averaging of results. The matrix M will therefore consist of the mean of the binary non-inferiority values $NI[G, St]$ throughout the range of potential St settings (y-axis) and throughout a range of EEG testing dataset sizes (x-axis).

$$M(Sn, nE) = \sum_1^z \text{mean} \left(\sum_1^{St} \sum_1^{nE} NI(G, St) \right) \quad (6.1)$$

If M is calculated moving from lower to higher values along the x-axis as increasing values of expert group size (n_i) are considered, there may or may not be a convergence towards a range for PTV (St) settings for the detector which provide reliable non-inferiority in performance (versus human experts) for the detector.

6.2.5.2 Investigating the effect of expert characteristics

Given that the number of groups of n size which can be chosen from within a group of 35 neurologists experts is $35!/(n!(35-n)!)$, and therefore incredibly large, a random sampling of 1000 groups of each of the expert group sizes between 3-12 was constructed. For each of these groups, we calculated IRA using Gwet's AC2; and also were calculated various features based on characteristics known to affect the performance of experts in labeling IEDs [150] including, (1) average number of years of EEG fellowship training, (2) fraction of the experts in the group with ABCN certification, (3) fraction of the experts in the group with ABPN-Epilepsy certification. A matrix P is defined; one column of matrix P can be calculated at $ni = nE$, as described above, and this column vector can be summed, as a measurement representing the degree to which the performance of the P13 (at all possible PTV settings) is non-inferior to the group of experts. Four 2D plots can be created with each element representing one of these summed columns within P with group size from 3-12 on the y-axis; and for each of the four plots, along the x-axis could be four characteristics of our neurologist expert groups which we have found to be relevant to IED IRA analysis [150] (1) the Gwet's AC2 IRA of each group, (2) the average years of EEG fellowship training, (3) the fraction of the group with ABCN certification, and (4) the fraction of the group with ABPN-E certification.

6.3 Results and Discussion

This section includes the results of the analysis on the performance of P13 spike detector.

6.3.1 Inter-rater agreement analysis

For each of the 19 academic EEGers, the average agreement with the remaining 18 experts (measured by Gwet's AC2) is shown in Fig. 6.3. For P13, mean agreement with the 19 experts is shown across the full range of PTVs (0.1 to 1) in Fig. 6.4. As seen in Fig. 6.4, the mean pairwise Gwet's AC2 of P13 with expert EEGers increases as PTV increases. The lowest mean pairwise Gwet's AC2 among the 19 experts is 0.5788, and P13 in $PTV \geq 0.8$ has a mean pairwise Gwet's AC2 greater than this. However, 90%

(17 of 19) of experts have an AC2 value between 0.84-0.90, while the maximum for P13, which occurs at $PTV = 1$, is 0.85. This PTV is 8 times higher than the value reported to match the IRA values of technologists in [137].

Figure 6.3: Mean pairwise Gwet's AC2 for all 19 annotators.

6.3.2 ROC based on experts' consensus

We compared the performance of P13 with every human expert by considering the consensus of differing numbers of experts. Since there are a large number of experts in this dataset, we computed the Se versus FPR considering consensus of at least 5, 8, 12, and 15 experts in Fig. 6.5 to Fig. 6.8, respectively. In these figures, FPR versus Se is plotted for the P13 algorithm at various thresholds, and also for human readers. We calculated P13's performance based on the same ground truth. Same color and style is utilized for P13 and experts, for all thresholds. In every plot, there are 18 curves corresponding to P13, where each curve shows the performance of P13 based on the consensus of the experts; on the same plot, the performance of each expert is displayed. Although some of these curves are overlapping, the plots provide a clear comparison between P13 and every individual expert, where different levels of expert agreement are considered. The figures show that

Figure 6.4: Mean pairwise Gwet's AC2 for P13 in thresholds from 0.1 to 1.

the performance of P13 based on consensus of experts is inferior to human annotators, and this difference becomes more pronounced for increasing number of experts in the consensus.

6.3.3 Mean pairwise ROC analysis

In Fig. 6.9, the pairwise Se versus FPR for P13 and all 19 human experts is illustrated. The performance of P13 detector at various perception thresholds is also plotted in this figure. Comparing mean pairwise performance for P13 versus the human experts, it is clear that for any given value for pairwise FPR, the pairwise Se of the P13 detector is lower than the experts.

6.3.4 Bootstrap statistical analysis of mean pairwise ROC

After performing BCa and finding the 95% confidence interval on all possible differences between pairwise values among all experts and the P13 detector, we calculated lower bound of sensitivity confidence interval and upper bound of FPR confidence interval. For accepting the hypothesis that an algorithm is non-inferior to an expert, the lowest lower

Figure 6.5: FPR versus sensitivity for P13 and human experts for sets of spikes marked by at least 5 experts. P13's performance is shown at the thresholds of 0.1 to 1. P13's and experts' performance is calculated based on the same ground truth and plotted with same color and style for all thresholds.

sensitivity difference bound for human expert pairs must be less than the lowest lower sensitivity difference bound of the algorithm versus human experts, and the highest upper FPR difference bound for the human expert pairs must be greater than the highest upper FPR difference bound of the algorithm versus expert bounds. According to the results of this study, the lower bound of sensitivity (Se) for human expert pairs was found to be $\delta_{sens} = -0.0668$ and for upper bound of FPR was calculated as $\delta_{FPR} = 2.23$ per minute.

The 0.95 confidence intervals for human expert pairs as well as P13-human pairs, for varying P13 thresholds are presented in Fig. 6.10. Two dotted lines show δ_{sens} and δ_{FPR} ; following the procedure of [137], if all 0.95 intervals for P13-human pairs for any threshold are above the horizontal (δ_{sens}) and to the left of the vertical line (δ_{FPR}), the algorithm is considered non-inferior to human experts at that threshold. According to Fig. 6.10, P13 satisfies this condition for the $PTS \geq 0.81$. In the study by Scheuer *et al.* [137], the hypothesis that P13 (at perception threshold setting of 0.1) meets non-inferiority criteria in comparison to a (tested) human reader was accepted. According

Figure 6.6: FPR versus sensitivity for P13 and human experts for sets of spikes marked by at least 8 experts. P13's performance is shown at the thresholds of 0.1 to 1. P13's and experts' performance is calculated based on the same ground truth, and plotted with same color and style for all thresholds.

to the analysis here, P13 passes this test when the perception threshold value (PTV) is greater than or equal 0.81, while according to [137], P13 passed the test with threshold equal to 0.1.

6.3.4.1 Investigating the effect of PTS and number of EEG subjects

We applied the non-inferiority test on subsets of EEG records of size $N_i = 2$ to 200 randomly sampled with replacement from the original 200 records. We considered the annotations of 19 clinical neurophysiologists. We applied 43 thresholds spaced logarithmically on P13 PTV. Each process was repeated $z = 30$ times and the average of all repetitions were computed as elements of M matrix, according to equation (6.1). Matrix M is illustrated in Fig. 6.11. This figure shows that for larger EEG datasets, at a significantly higher threshold of approximately 0.96, P13 could have a performance comparable to human experts.

Figure 6.7: FPR versus sensitivity for P13 and human experts for sets of spikes marked by at least 12 experts. P13's performance is shown at the thresholds of 0.1 to 1. P13's and experts' performance is calculated based on the same ground truth, and plotted with same color and style for all thresholds.

6.3.4.2 Investigating the effect of expert characteristics

Next, in order to investigate the effects of EEGer characteristics, we applied the non-inferiority test on the annotations from all 35 experts by considering subsets experts with group size 3 to 12, and 1000 possible groups from each group size. We applied the same 43 thresholds on the P13 PTV. The *NI* values for each group size averaged for all thresholds, versus 4 EEGer characteristics: (1) the Gwet's AC2 IRA of each group, (2) the fraction of the group with ABCN certification, (3) the fraction of the group with ABPN-E certification, and (4) the average years of EEG fellowship training; and are shown in Fig. 6.12, Fig. 6.13, Fig. 6.14, Fig. 6.15, respectively. In all of these plots, we observe that using a greater number of experts, as well as higher value for AC2, fraction of ABCN certification, fraction of ABPN-E certification, and the average years of EEG fellowship training yields higher *NI* values. This finding implies the importance of size and quality of expert group used for annotating and validation of IED detection software.

Figure 6.8: FPR versus sensitivity for the P13 algorithm and human readers for P13 and human experts for sets of spikes marked by at least 15 experts. P13's performance is shown at the thresholds of 0.1 to 1. P13's and experts' performance is calculated based on the same ground truth, and plotted with same color and style for all thresholds.

Figure 6.9: Average pairwise performance for P13 at different thresholds and human experts. P13's performance is calculated at the thresholds of 0.1 to 1.

Figure 6.10: Accelerated bootstrap performance test on the sensitivity and FPR values for P13's perception threshold varied from 0.1 to 1.

Figure 6.11: Non-inferiority test outcome applied on EEG datasets of size 2 to 200 at 43 PTVs for P13 smoothed by 30 repetitions.

Figure 6.12: Non-inferiority test outcome for scorer groups of size 3 to 12 averaged over all 43 PTVs, versus the Gwet's AC2 IRA of each group. (b) the average years of EEG fellowship training, (c) the fraction of the group with ABCN certification, and (d) the fraction of the group with ABPN-E certification.

Figure 6.13: Non-inferiority test outcome for scorer groups of size 3 to 12 averaged over all 43 PTVs, versus the fraction of the group with ABCN certification

Figure 6.14: Non-inferiority test outcome for scorer groups of size 3 to 12 averaged over all 43 PTVs, versus the fraction of the group with ABPN-E certification.

Figure 6.15: Non-inferiority test outcome for scorer groups of size 3 to 12 averaged over all 43 PTVs, versus the average years of EEG fellowship training.

6.4 Conclusions

In this study, we developed and implemented several tests to assess the performance of an IED detection system. We compared the P13 algorithm to 19 academic clinical neurophysiologists on a dataset consisting of EEG segments with and without IEDs of 200 patients. The mean pairwise sensitivity (Se) and false positive rate (FPR) performance of the P13 algorithm versus 19 academic EEGers showed that experts' performance is superior to P13 at all perception threshold settings.

We performed a test similar to what is referred to as "statistical Turing test" in [137] using bias-corrected and accelerated bootstrap analysis of the pairwise differences of sensitivity and FPR for each EEG records. Results showed that the P13 algorithm provides a level of performance equivalent to human experts at perception threshold setting of greater than 0.8 only, which is a different threshold reported by [137]. It is unclear whether this is the proper setting for the P13 perception threshold or if the ideal perception threshold setting varies depending on the EEG dataset.

Similar findings were obtained from inter-rater agreement analysis using Gwet AC2, which showed that the performance of P13 algorithm is comparable to experts at the perception threshold setting of greater or equal to 0.8.

We also investigated the effect of number of EEG subjects by utilizing the same test. The results showed that for larger EEG datasets, at a significantly higher threshold of approximately 0.96, P13 could have a performance comparable to human experts. In addition, we investigated the effect of expert characteristics by applying the non-inferiority test. We observed that using more number of experts, as well as higher value of inter-rates agreement among them, higher fraction of ABCN certification, higher fraction of ABPN-E certification, and more years of EEG fellowship training results in higher *NI* values. This shows the importance of size and quality of expert group considered for annotation and validation of IED detection software.

Experts agree that EEGs are frequently misinterpreted due to misidentification of IEDs [191]. There is a need for accurate IED detection systems to assist neurologists with EEG interpretation. Unfortunately, the analysis performed in this study suggests that the P13 algorithm's IED detection performance is not equivalent to that of the average aca-

demic clinical neurophysiologist. This study also shows that the default P13 perception threshold setting is inappropriately set to 0.1, since we observed that the P13 algorithm performs considerably better at a threshold set to higher values. Since this study was performed on only 200 brief segments of EEG, further studies need to be performed on a larger number of longer EEG recordings in order to confirm our findings. Even though based on this analysis, P13 can have a performance comparable to human experts at a higher threshold than the default, there are several limitations for this approach. Firstly, it is not clear a priori that what PTS would perform well on a new dataset. This probably has to be assessed using more EEG datasets. Secondly, there are several issues with the non-inferiority test suggested by Scheuer *et al.* In this approach all spike annotations are considered statistically independent, even though they are not. There are a number of statistical dependencies in the dataset including:

- Dependency within an EEG recording due to level of epileptogenic appearance of the events within that EEG recording: Since some EEG datasets just have IEDs with either a higher than average or lower than average mean rating for their spikeness among all scorers, so that these datasets will potentially contribute more to effect size. These outlier datasets would include datasets with very obvious IEDs which are classic in appearance, or EEG datasets with more subtle IEDs (lower amplitude, no after-going slow wave, broad based sharp waves, etc.)
- Dependency of spike frequency within an EEG recording: As some EEG recordings have a higher temporal density of IEDs, or the EEG recordings are longer, meaning that such an EEG dataset could contribute more data points to the non-inferiority statistical calculation, in the way that this method has been set up.
- Dependency within a single scorer's rating data: Since some EEGers mark either a lot more IEDs at average throughout all EEG datasets, than the average scorer, or mark much fewer IEDs in all EEG datasets at average, than other EEG scorers.

Moreover, in this NI approach, the non-inferiority region was defined so that all experts would be considered non-inferior to each other. This is problematic if one or more of the scorers is unskilled or has poor IRA with the others. Including the less skilled scorers substantially lowers the bar for the algorithm. Therefore, passing the NI test while

considering all possible experts to evaluate an algorithm, can mean that the algorithm is only comparable to the poorest EEGers in the dataset. This is because the NI threshold is too dependent on the poorest scorer, and the dependency between all the false positive annotations by a poor scorer within a particular EEG recording is not accounted for in the statistical approach. More has to be done on this aspect in order to conduct a NI test using only skilled and experienced EEGers and avoid outliers.

Chapter 7

Conclusion and Recommendations

In this chapter, we present a summary of this thesis' key results, advantages and limitations. We also propose Some potential ideas and methods to be applied in the future work.

7.1 Background Rejection Algorithm

In order to design an efficient system for IED detection, we proposed an ensemble method to perform fast multi-step background rejection on interictal EEG of patients with epilepsy. We developed a huge feature pool comprising several features such as morphological features, energy, continuous and discrete wavelet analyzes, nonlinear energy operator, and spectrogram features. By applying thresholds on these features consecutively, we formed a cascade of thresholding steps which was then applied to reject partial background waveforms in each stage.

Using a sufficiently large dataset consisting of several different subjects including epileptic patients as well as IED-free subjects ensures that the cascade can generalize quite well for new EEG recordings of different signal to noise ratio. In addition, we applied cross-validation on the dataset.

Results showed that by applying this algorithm, we are able to reject a large proportion of background waveforms and maintain most of the IEDs. Specifically, this method rejects 78% of the background waveforms while retaining 97% of the IEDs, as

seen in Table 3.2 and Table 3.3. Next, we compared two classification scenarios: firstly, we applied classifiers such as support vector machine, k nearest neighbor, and random forests to classify EEG; secondly, the we applied the classifiers after rejecting the background from EEG. By comparing the performance and speed of these two methods, we observed that background rejection speeds up IED detection by eliminating most unwanted EEG portions at an initial phase. The speed-up ratio was ranging from 4.69 to 1.76 as shown in Table 3.5. Since speed is a crucial factor for an IED detection system, we concluded that by utilizing this algorithm, we can save time while applying more complicated pattern recognition methods to further classify the IEDs in large EEG datasets. By filtering out a large proportion of background data by this proposed algorithm, a reasonable computational complexity can be maintained for the overall IED detection system.

In addition, it was shown that this algorithm can be regarded as a feature selection technique as well; and the feature ranking obtained in each stage, revealed the most discriminative features which can be employed to separate IEDs from background EEG activity.

One limitation of this method is that it does not improve the performance of IED detection. Rather, it only lowers the computation time. While this is the main goal of developing this algorithm, it is meaningful to make modifications in such a way that we can improve results in terms of sensitivity and false positive rate as well.

A potential way for this purpose is to apply other feature selection methods to find the feature ranking. So far, CDF value was utilized to select features for each step. Exploring other methods might change the feature order, and possibly improve the results. Using faster feature selection methods can also help us decrease the time needed to build the cascade.

Moreover, we can increase the threshold, such that fewer IEDs would be lost in each stage. By setting higher thresholds and leveraging on more diverse EEG features, we may increase length of the cascade, and achieve higher sensitivity and specificity. This requires an even larger dataset of IED and background EEG data to avoid overfitting. In addition, applying the method on an even larger number of subjects will provide more reliable performance characteristics of this method in general.

7.2 Classifier Cascade

We designed a cascade of simple classifiers (SVMs) in which each classifier was trained by different sets of background waveforms. This algorithm was based on rejecting background EEG activity from interictal EEG of patients with epilepsy using their ROC, with the purpose of improving the performance of IED detection systems in terms of precision and false positive rate. We considered a large dataset consisting of 93 epileptic subjects as well as 63 subjects with IED-free EEG. Various sizes can be selected for the cascade as well, to pick the optimum for our application. Our results in Fig. 4.4 and Fig. 4.3 showed that by applying the SVM cascade, at a fixed sensitivity the IED detection precision improves by 6.78%; and at a fixed false positive rate the sensitivity improves by 2.83%.

In the future work, we can employ other types of classifiers as the basic units of the algorithm. SVM was applied in this study; however, we can include classifiers such as KNN, LDA, ELM, etc.; this can further improve precision and false positive rate. We can employ different types of classifiers at the same time, as well. By doing so, we can find the optimum set of classifier sequence to be applied for IED detection. Moreover, this method can be applied on features rather than waveform time series data. Computationally efficient feature selection methods can be applied to utilize the selected features as the input to the algorithm. We can also investigate the algorithm with more IED detection methods.

Last but not the least, IEDs that are lost at the early stages of the cascade can also be analyzed. In this way, we can learn more about morphologies of IEDs that can easily be misinterpreted by IED detection algorithms. Furthermore, analyzing the background waveforms which are retained after the later stages of the algorithm can be useful. These waveforms comprise difficult-to-reject false positive detections; therefore, studying their morphologies and trying to eliminate them could yield further reduction in the false positive rate of IED detection methods in the future.

7.3 Inter-rater Agreement Analysis in IED Detection

We analyzed EEG events annotated by a large number of experts. We investigated inter-rater agreement among the EEGers, extracted several features for each waveform and analyzed them in terms of their relationship with expert scores assigned to each event.

We evaluated numerous EEG features to determine which ones best predict whether a waveform will be judged by experts to be an IED, and which features best predict differences of opinion. We found that expert opinion regarding IEDs is predictable when the number of experts is large.

By means of correlation analysis and elastic net feature selection, we proposed a set of features which can be beneficial in characterizing the scores that experts give to an event, and the variance in scores given by different experts. The key features were presented in Table 5.2, Table 5.2, Fig. 5.16 and Fig. 5.17. Wavelet coefficients were observed to be helpful in characterizing most of the different aspects of IEDs and therefore, beneficial in automated IED detection. We employed T-SNE technique to visualize waveforms in two dimensions, giving good discrimination between events with different scores, using the extracted features.

The FDA currently only requires three EEGers to annotate the location of IEDs in a testing dataset for approval of an IED detection algorithm. Nevertheless, the results of the present study including the RMSE values shown in Fig. 5.19 suggest that using more than three EEGers will probably produce better results.

In addition, we applied support vector regression to predict the scores that experts assign to an event in terms of being an IED. We were able to predict average scores of experts with relatively low error values, using the selected features. The prediction ROC and precision-recall curves were presented in Fig. 5.22 and Fig. 5.22. We observed better prediction is achieved if normalization is performed by locally standardized EEG by z-scores. Globally normalizing each channel of EEG data did not improve the prediction error. We presented the numerical results in Table 5.5. This is a useful finding since it may inform best practices regarding preprocessing steps for future IED detection algorithms.

We also analyzed the relation of electrode locations in which events are annotated

with the scores experts gave to each event (see Table 5.4). We found locations on the scalp, which EEGers annotate IEDs with higher confidence.

In summary, the results of this study show that expert opinion regarding agreement or disagreement in the interpretation of IEDs is predictable. This analysis also identified preprocessing steps and particular features which are effective for training machine learning algorithms to predict expert agreement regarding IEDs, including local normalization of waveforms, and particular types of wavelet functions, among other features. In addition, this analysis reveals that aggregated scores from a large number of experts are substantially more predictable than scores from small numbers of experts. This finding suggests that despite studies finding that inter-rater agreement among experts is imperfect, there is an underlying objective and predictable reality that can be captured by machine learning algorithms, and also the common practice of seeking consensus by group review on difficult EEG cases may be helpful in getting closer to the truth in detecting IEDs.

Since our ultimate goal is to develop efficient IED detection software, with very low number of false detections, and high accuracy, findings from this study give us better understanding of EEG datasets and features, which can be employed for IED detection. Having the features, which correlate with expert scores in different aspects, as well as finding out that using scores of several scorers leads to more accurate predictions rather than using only selected scorers, are examples that can help us develop more reliable algorithms for IED detection.

One limitation of this study is using a dataset of short EEG segments; the number of EEG events considered is probably lower than the number necessary to completely encompass the variability found in epileptiform and benign epileptiform-appearing signals in all subjects. Moreover, IRA between experts might have been higher if the EEG segments from each subject had been longer than 30 seconds. Future studies will need to exploit larger EEG datasets and more subjects.

The dataset also has a high degree of marginal inhomogeneity in the expert opinion; this can be reduced by providing more detailed instructions to experts while scoring the EEG.

7.4 Computers versus Humans

We proposed and implemented several methods to analyze and test the performance of an IED detection system. We tested and compared the P13 algorithm to 35 experts including 19 academic clinical neurophysiologists on a dataset consisting of EEG segments with and without IEDs of 200 patients.

This algorithm is claimed to be non-inferior to humans at the perception value of 0.1. The results of this study showed that experts' performance is superior to P13 at all perception threshold settings in terms of the pairwise sensitivity versus false positive rate as shown in Fig. 6.9. Moreover, performing a test similar to what is referred to as "statistical Turing test" in [137] which we refer to as non-inferiority test, showed that the P13 algorithm provides a level of performance equivalent to human experts at a different threshold reported by [137]. The outcome at various threshold settings was presented in Fig. 6.10. It is unclear whether this is the proper setting for the P13 perception threshold or if the ideal perception threshold setting varies depending on the EEG dataset. Similar findings were obtained from inter-rater agreement analysis.

We also investigated the effect of EEG dataset size and a number of expert characteristics on the outcome of the non-inferiority test. We determined the importance of dataset size and quality of expert group considered for annotation and validation of IED detection systems.

Experts agree that EEGs are frequently misinterpreted due to misidentification of IEDs, hence, there is a need for accurate IED detection systems to assist neurologists with EEG interpretation. Unfortunately, the analysis performed in this study suggests that the P13 algorithm's IED detection performance is not equivalent to that of the average academic clinical neurophysiologist, and the default P13 perception threshold setting is inappropriately set to 0.1.

The present study was performed on only 200 brief segments of EEG, and further studies need to be performed on a larger number and longer EEG recordings in order to confirm its findings.

We provided a comprehensive group of statistical tests and measures to evaluate an IED detection system. By minor modifications these methods can be applied on other

types of algorithms.

considering the limitations explained in Chapter 6, a modified non-inferiority test which takes into account the dependencies in the dataset, and expert outliers, would help drawing more realistic conclusions from a NI test on the performance of a detection algorithm. This chapter provided a summary of main conclusions made from the work performed in this thesis on IED detection. We tackled IED detection in different aspects from improvement in speed, and reducing false detections, to analyzing how experts interpret interictal EEG, and how automated software needs to be properly validated for IED detection. It also included recommendations to improve the existing work for future research in this field. Detailed methods and results can be found in Chapters 3 to 6.

Author's Publications

Journal Papers

1. E. Bagheri, J. Dauwels, B. C. Dean, C. G. Waters, M. B. Westover, and J. J. Halford, "Interictal epileptiform discharge characteristics underlying expert interrater agreement," *Clinical Neurophysiology*, vol. 128, no. 10, pp. 1994-2005, 2017.
2. E. Bagheri, J. Jin, J. Dauwels, S. Cash, and M. B. Westover, "A fast and automated method to facilitate the detection of interictal epileptiform discharges in the scalp electroencephalogram," Under review by *Journal of Neuroscience Methods*.

Conference Papers

1. E. Bagheri, J. Jin, J. Dauwels, S. Cash, and M. B. Westover, "Fast and efficient rejection of background waveforms in interictal EEG," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 744-748.
2. E. Bagheri, J. Jin, J. Dauwels, S. Cash, and M. B. Westover, "Classifier cascade to aid in detection of epileptiform transients in interictal EEG," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 970-974.
3. E. Bagheri, J. Dauwels, B. C. Dean, C. G. Waters, M. B. Westover, and J. J. Halford, "Epileptiform transient characteristics underlying expert interrater agreement," *Journal of Clinical Neurophysiology, Poster Abstract Supplement, ACNS 2017 Annual Meeting & Courses*, pp. 20, 2017.

Bibliography

- [1] K.N. Simpson , Unpublished data from medicare 5% sample and Truven Marketscan 2013 data set. Personal communication.
- [2] G. Chatrian, L. Bergamini, M. Dondey, D. Klass, M. Lennox-Buchthal, and I. Petersen, “A glossary of terms most commonly used by clinical electroencephalographers,” *Electroencephalogr Clin Neurophysiol*, vol. 37, no. 5, pp. 538–548, 1974.
- [3] J. J. Halford, “Computerized epileptiform transient detection in the scalp electroencephalogram: Obstacles to progress and the example of computerized {ECG} interpretation,” *Clinical Neurophysiology*, vol. 120, no. 11, pp. 1909 – 1915, 2009. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1388245709004921>
- [4] C. A. van Donselaar, R.-J. Schimsheimer, A. T. Geerts, and A. C. Declerck, “Value of the electroencephalogram in adult patients with untreated idiopathic first seizures,” *Archives of neurology*, vol. 49, no. 3, pp. 231–237, 1992.
- [5] S. Seidel, E. Pablik, S. Aull-Watschinger, B. Seidl, and E. Patariaia, “Incidental epileptiform discharges in patients of a tertiary centre,” *Clinical Neurophysiology*, vol. 127, no. 1, pp. 102–107, 2016.
- [6] N. B. Fountain and J. M. Freeman, “EEG is an essential clinical tool: pro and con,” *Epilepsia*, vol. 47, no. s1, pp. 23–25, 2006.
- [7] J. J. Halford, R. J. Schalkoff, J. Zhou, S. R. Benbadis, W. O. Tatum, R. P. Turner, S. R. Sinha, N. B. Fountain, A. Arain, P. B. Pritchard *et al.*, “Standardized database development for EEG epileptiform transient detection: EEGnet scoring system

- and machine learning analysis,” *Journal of neuroscience methods*, vol. 212, no. 2, pp. 308–316, 2013.
- [8] S. R. Benbadis, “Errors in EEGs and the misdiagnosis of epilepsy: importance, causes, consequences, and proposed remedies,” *Epilepsy & Behavior*, vol. 11, no. 3, pp. 257–262, 2007.
- [9] T. M. Dall, M. V. Storm, R. Chakrabarti, O. Drogan, C. M. Keran, P. D. Donofrio, V. W. Henderson, H. J. Kaminski, J. C. Stevens, and T. R. Vidic, “Supply and demand analysis of the current and future us neurology workforce,” *Neurology*, vol. 81, no. 5, pp. 470–478, 2013.
- [10] M. Reuber, G. Fernandez, J. Bauer, C. Helmstaedter, and C. E. Elger, “Diagnostic delay in psychogenic nonepileptic seizures,” *Neurology*, vol. 58, no. 3, pp. 493–495, 2002.
- [11] L. Ver Hoef, R. Elgavish, and R. C. Knowlton, “Effect of detection parameters on automated electroencephalography spike detection sensitivity and false-positive rate,” *Journal of Clinical Neurophysiology*, vol. 27, no. 1, pp. 12–16, 2010.
- [12] M. De Lucia, J. Fritschy, P. Dayan, and D. S. Holder, “A novel method for automated classification of epileptiform activity in the human electroencephalogram-based on independent component analysis,” *Medical & biological engineering & computing*, vol. 46, no. 3, pp. 263–272, 2008.
- [13] K. Indiradevi, E. Elias, P. Sathidevi, S. D. Nayak, and K. Radhakrishnan, “A multi-level wavelet approach for automatic detection of epileptic spikes in the electroencephalogram,” *Computers in biology and medicine*, vol. 38, no. 7, pp. 805–816, 2008.
- [14] T. E. Losey and L. Uber-Zak, “Time to first interictal epileptiform discharge in extended recording EEGs,” *Journal of Clinical Neurophysiology*, vol. 25, no. 6, pp. 357–360, 2008.

- [15] P. N. Modur and B. Rigdon, "Diagnostic yield of sequential routine EEG and extended outpatient video-EEG monitoring," *Clinical Neurophysiology*, vol. 119, no. 1, pp. 190–196, 2008.
- [16] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [17] R. S. Fisher, C. Acevedo, A. Arzimanoglou, A. Bogacz, J. H. Cross, C. E. Elger, J. Engel Jr, L. Forsgren, J. A. French, M. Glynn *et al.*, "Ilae official report: a practical clinical definition of epilepsy," *Epilepsia*, vol. 55, no. 4, pp. 475–482, 2014.
- [18] W. H. Organization. (2018) Epilepsy. [Online]. Available: <http://www.who.int/news-room/fact-sheets/detail/epilepsy>
- [19] M. Pandolfo, "Genetics of epilepsy," in *Seminars in neurology*, vol. 31, no. 05. © Thieme Medical Publishers, 2011, pp. 506–518.
- [20] R. S. Fisher, W. V. E. Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel Jr, "Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe)," *Epilepsia*, vol. 46, no. 4, pp. 470–472, 2005.
- [21] E. Foundation. (2018) What is epilepsy? [Online]. Available: <https://www.epilepsy.com/learn/about-epilepsy-basics/what-epilepsy>
- [22] I. L. A. Epilepsy. (2018) Guidelines. [Online]. Available: <https://www.ilae.org/guidelines>
- [23] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.
- [24] G. E. of Surgery: A Guide for Patients and Caregivers. (2018) Electroencephalography. [Online]. Available: <https://www.encyclopedia.com/medicine/divisions-diagnostics-and-procedures/medicine/electroencephalography>

- [25] W. O. Tatum IV, *Handbook of EEG interpretation*. Demos Medical Publishing, 2014.
- [26] C. C. Chernecky and B. J. Berger, *Laboratory tests and diagnostic procedures*. Elsevier Health Sciences, 2007.
- [27] I. F. of Clinical Neurophysiology. (2018) Guidelines. [Online]. Available: <http://www.ifcn.info/showcontent.aspx?MenuID=1169>
- [28] S. Noachtar, C. Binnie, J. Ebersole, F. Mauguiere, A. Sakamoto, and B. Westmoreland, “A glossary of terms most commonly used by clinical electroencephalographers and proposal for the report form for the EEG findings. the international federation of clinical neurophysiology.” *Electroencephalography and clinical neurophysiology. Supplement*, vol. 52, p. 21, 1999.
- [29] S. B. Wilson and R. Emerson, “Spike detection: a review and comparison of algorithms,” *Clinical Neurophysiology*, vol. 113, no. 12, pp. 1873–1881, 2002.
- [30] S. BÖLGEN and G. KÖKTÜRK, “EEG spike classification with template matching algorithm.”
- [31] J. R. Stevens, B. L. Lonsbury, and S. L. Goel, “Seizure occurrence and inter-spike interval: telemetered electroencephalogram studies,” *Archives of neurology*, vol. 26, no. 5, p. 409, 1972.
- [32] G. Fischer, N. Mars, and F. L. da Silva, “Pattern recognition of epileptiform transients in the electroencephalogram,” *Da Costakade, Utrecht, The Netherlands: Institute of Medical Physics*, 1980.
- [33] S. S. Lodder, J. Askamp, and M. J. van Putten, “Inter-ictal spike detection using a database of smart templates,” *Clinical Neurophysiology*, vol. 124, no. 12, pp. 2328 – 2335, 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1388245713006895>
- [34] A. Nonclercq, M. Foulon, D. Verheulpen, C. De Cock, M. Buzatu, P. Mathys, and P. Van Bogaert, “Cluster-based spike detection algorithm adapts to interpatient and

- inpatient variation in spike morphology,” *Journal of neuroscience methods*, vol. 210, no. 2, pp. 259–265, 2012.
- [35] J. Jin, J. Dauwels, S. Cash, and M. Westover, “Spikegui: Software for rapid interictal discharge annotation via template matching and online machine learning,” in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, Aug 2014, pp. 4435–4438.
- [36] P. Stoica and R. L. Moses, *Spectral analysis of signals*. Pearson/Prentice Hall Upper Saddle River, NJ, 2005.
- [37] K. Arakawa, D. H. Fender, H. Harashima, H. Miyakawa, and Y. Saitoh, “A nonlinear digital filter for separating nonstationary waves from stationary waves of the EEG,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’86.*, vol. 11. IEEE, 1986, pp. 1809–1812.
- [38] F. Lopes da Silva, A. Dijk, and H. Smits, “Detection of non-stationarities in EEGs using the autoregressive model—an application to EEGs of epileptics, in computerized EEG analysis,” 1975.
- [39] G. Pfurtscheller and G. Fischer, “A new approach to spike detection using a combination of inverse and matched filter techniques,” *Electroencephalography and clinical neurophysiology*, vol. 44, no. 2, pp. 243–247, 1978.
- [40] H. S. Liu, T. Zhang, and F. S. Yang, “A multistage, multimethod approach for automatic detection and classification of epileptiform EEG,” *Biomedical Engineering, IEEE Transactions on*, vol. 49, no. 12, pp. 1557–1566, 2002.
- [41] R. Sankar and J. Natour, “Automatic computer analysis of transients in EEG,” *Computers in biology and medicine*, vol. 22, no. 6, pp. 407–422, 1992.
- [42] L. Senhadji, J.-L. Dillenseger, F. Wendling, C. Rocha, and A. Kinie, “Wavelet analysis of EEG for three-dimensional mapping of epileptic events,” *Annals of biomedical engineering*, vol. 23, no. 5, pp. 543–552, 1995.

- [43] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
- [44] A. A. Dingle, R. D. Jones, G. J. Carroll, and W. R. Fright, "A multistage system to detect epileptiform activity in the EEG," *Biomedical Engineering, IEEE Transactions on*, vol. 40, no. 12, pp. 1260–1268, 1993.
- [45] W. R. S. Webber, B. Litt, R. Lesser, R. Fisher, and I. Bankman, "Automatic EEG spike detection: what should the computer imitate?" *Electroencephalography and clinical neurophysiology*, vol. 87, no. 6, pp. 364–373, 1993.
- [46] P. G. De Oliveira, C. Queiroz, and F. L. Da Silva, "Spike detection based on a pattern recognition approach using a microcomputer," *Electroencephalography and clinical neurophysiology*, vol. 56, no. 1, pp. 97–103, 1983.
- [47] J. Gotman, J. Ives, and P. Gloor, "Automatic recognition of inter-ictal epileptic activity in prolonged EEG recordings," *Electroencephalography and clinical neurophysiology*, vol. 46, no. 5, pp. 510–520, 1979.
- [48] W. HOSTETLER, H. DOLLER, and R. HOMAN, "ASSESSMENT OF A COMPUTER-PROGRAM TO DETECT EPILEPTIFORM SPIKES," *ELECTROENCEPHALOGRAPHY AND CLINICAL NEUROPHYSIOLOGY*, vol. 83, no. 1, pp. 1–11, JUL 1992.
- [49] T. Pietilä, S. Vapaakoski, U. Nousiainen, A. Värri, H. Frey, V. Häkkinen, and Y. Neuvo, "Evaluation of a computerized system for recognition of epileptic activity during long-term EEG recording," *Electroencephalography and clinical neurophysiology*, vol. 90, no. 6, pp. 438–443, 1994.
- [50] R. Benlamri, M. Batouche, S. Rami, and C. Bouanaka, "An automated system for analysis and interpretation of epileptiform activity in the EEG," *Computers in biology and medicine*, vol. 27, no. 2, pp. 129–139, 1997.
- [51] T. P. Exarchos, A. T. Tzallas, D. I. Fotiadis, S. Konitsiotis, and S. Giannopoulos, "EEG transient event detection and classification using association rules," *In-*

- formation Technology in Biomedicine, IEEE Transactions on*, vol. 10, no. 3, pp. 451–457, 2006.
- [52] P. Y. Ktonas and J. R. Smith, “Quantification of abnormal eeg spike characteristics,” *Computers in biology and medicine*, vol. 4, no. 2, pp. 157–163, 1974.
 - [53] C. J. James, R. D. Jones, P. J. Bones, and G. J. Carroll, “Detection of epileptiform discharges in the EEG by a hybrid system comprising mimetic, self-organized artificial neural network, and fuzzy logic stages,” *Clinical Neurophysiology*, vol. 110, no. 12, pp. 2049–2063, 1999.
 - [54] M. Adjouadi, M. Cabrerizo, M. Ayala, D. Sanchez, I. Yaylali, P. Jayakar, and A. Barreto, “A new mathematical approach based on orthogonal operators for the detection of interictal spikes in epileptogenic data,” *Biomedical sciences instrumentation*, vol. 40, pp. 175–180, 2003.
 - [55] M. Adjouadi, D. Sanchez, M. Cabrerizo, M. Ayala, P. Jayakar, I. Yaylali, and A. Barreto, “Interictal spike detection using the walsh transform,” *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 5, pp. 868–872, 2004.
 - [56] T. Sugi, M. Nakamura, A. Ikeda, and H. Shibasaki, “Adaptive EEG spike detection: determination of threshold values based on conditional probability,” *Frontiers of Medical and Biological Engineering*, vol. 11, no. 4, pp. 261–277, 2001.
 - [57] J. R. Glover Jr, N. Raghaven, P. Y. Ktonas, and J. D. Frost Jr, “Context-based automated detection of epileptogenic sharp transients in the EEG: elimination of false positives,” *Biomedical Engineering, IEEE Transactions on*, vol. 36, no. 5, pp. 519–527, 1989.
 - [58] M. A. Black, R. D. Jones, G. J. Carroll, A. A. Dingle, I. M. Donaldson, and P. J. Parkin, “Real-time detection of epileptiform activity in the EEG: a blinded clinical trial,” *Clinical EEG and Neuroscience*, vol. 31, no. 3, pp. 122–130, 2000.
 - [59] J. Gotman and L. Wang, “State-dependent spike detection: concepts and preliminary results,” *Electroencephalography and clinical Neurophysiology*, vol. 79, no. 1, pp. 11–19, 1991.

- [60] J. Gotman and L.-Y. Wang, "State dependent spike detection: validation," *Electroencephalography and clinical neurophysiology*, vol. 83, no. 1, pp. 12–18, 1992.
- [61] Y.-C. Liu, C.-C. K. Lin, J.-J. Tsai, and Y.-N. Sun, "Model-based spike detection of epileptic EEG data," *Sensors*, vol. 13, no. 9, pp. 12 536–12 547, 2013.
- [62] N. Acır, I. Öztura, M. Kuntalp, B. Baklan, and C. Güzelis, "Automatic detection of epileptiform events in EEG by a three-stage procedure based on artificial neural networks," *Biomedical Engineering, IEEE Transactions on*, vol. 52, no. 1, pp. 30–40, 2005.
- [63] J. S. Barlow, "EEG transient detection by matched inverse digital filtering," *Electroencephalography and clinical neurophysiology*, vol. 48, no. 2, pp. 246–248, 1980.
- [64] M. Feucht, K. Hoffmann, K. Steinberger, H. Witte, F. Benninger, M. Arnold, and A. Doering, "Simultaneous spike detection and topographic classification in pediatric surface EEGs," *NeuroReport*, vol. 8, no. 9, pp. 2193–2197, 1997.
- [65] A. V. Oppenheim, R. W. Schaffer, J. R. Buck *et al.*, *Discrete-time signal processing*. Prentice hall Englewood Cliffs, NJ, 1989, vol. 2.
- [66] S. L. Miller and D. Childers, "{CHAPTER} 8 - random processes," in *Probability and Random Processes (Second Edition)*, second edition ed., S. L. M. Childers, Ed. Boston: Academic Press, 2012, pp. 335 – 382. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/B9780123869814500114>
- [67] M. Latka, Z. Was, A. Kozik, and B. J. West, "Wavelet analysis of epileptic spikes," *Physical Review E*, vol. 67, no. 5, p. 052902, 2003.
- [68] I. Güler and E. D. Übeyli, "Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients," *Journal of neuroscience methods*, vol. 148, no. 2, pp. 113–121, 2005.
- [69] S. Mallat, *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

- [70] A. Graps, "An introduction to wavelets," *IEEE computational science and engineering*, vol. 2, no. 2, pp. 50–61, 1995.
- [71] R. Eberhart, R. Dobbins, and W. Webber, "EEG waveform analysis using casenet," in *Engineering in Medicine and Biology Society, 1989. Images of the Twenty-First Century., Proceedings of the Annual International Conference of the IEEE Engineering in.* IEEE, 1989, pp. 2046–2047.
- [72] Ö. Özdamar and T. Kalayci, "Detection of spikes with artificial neural networks using raw EEG," *Computers and Biomedical Research*, vol. 31, no. 2, pp. 122–142, 1998.
- [73] C.-W. Ko and H.-W. Chung, "Automatic spike detection via an artificial neural network using raw EEG data: effects of data preparation and implications in the limitations of online recognition," *Clinical Neurophysiology*, vol. 111, no. 3, pp. 477–481, 2000.
- [74] C. Kurth, F. Gilliam, and B. Steinhoff, "EEG spike detection with a kohonen feature map," *Annals of Biomedical Engineering*, vol. 28, no. 11, pp. 1362–1369, 2000.
- [75] S. B. Wilson, C. A. Turner, R. G. Emerson, and M. L. Scheuer, "Spike detection ii: automatic, perception-based detection and clustering," *Clinical neurophysiology*, vol. 110, no. 3, pp. 404–411, 1999.
- [76] W. Webber, B. Litt, K. Wilson, and R. Lesser, "Practical detection of epileptiform discharges (eds) in the EEG using an artificial neural network: a comparison of raw and parameterized EEG data," *Electroencephalography and clinical neurophysiology*, vol. 91, no. 3, pp. 194–204, 1994.
- [77] A. Tzallas, P. Karvelis, C. Katsis, D. Fotiadis, S. Giannopoulos, S. Konitsiotis *et al.*, "A method for classification of transient events in EEG recordings: application to epilepsy diagnosis," *Methods of Information in Medicine*, vol. 45, no. 6, pp. 610–621, 2006.

- [78] A. J. Gabor and M. Seyal, "Automated interictal EEG spike detection using artificial neural networks," *Electroencephalography and Clinical neurophysiology*, vol. 83, no. 5, pp. 271–280, 1992.
- [79] C. Castellaro, G. Favaro, A. Castellaro, A. Casagrande, S. Castellaro, D. Puthenparampil, and C. F. Salimbeni, "An artificial intelligence approach to classify and analyse EEG traces," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 32, no. 3, pp. 193–214, 2002.
- [80] T. Kalayci and Ö. Özdamar, "Wavelet preprocessing for automated neural network detection of EEG spikes," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 14, no. 2, pp. 160–166, 1995.
- [81] F. I. Argoud, F. M. De Azevedo, J. M. Neto, and E. Grillo, "Sade3: an effective system for automated detection of epileptiform events in long-term EEG based on context information," *Medical and Biological Engineering and Computing*, vol. 44, no. 6, pp. 459–470, 2006.
- [82] H. S. Park, Y. H. Lee, N. G. Kim, D. S. Lee, and S. I. Kim, "Detection of epileptiform activities in the EEG using neural network and expert system." *Studies in health technology and informatics*, vol. 52, pp. 1255–1259, 1997.
- [83] J. R. Glover Jr, D. N. Varmazis, P. Y. Ktonas, and J. D. Frost Jr, "Continued development of a knowledge-based system to detect epileptogenic sharp transients in the EEG," in *Engineering in Medicine and Biology Society, 1990., Proceedings of the Twelfth Annual International Conference of the IEEE.* IEEE, 1990, pp. 1374–1375.
- [84] B. Davey, W. Fright, G. Carroll, and R. Jones, "Expert system approach to detection of epileptiform activity in the EEG," *Medical and Biological Engineering and Computing*, vol. 27, no. 4, pp. 365–370, 1989.
- [85] H. J. Carey, M. Manic, and P. Arsenovic, "Epileptic spike detection with EEG using artificial neural networks," in *Human System Interactions (HSI), 2016 9th International Conference on.* IEEE, 2016, pp. 89–95.

- [86] D. H. Hubel and T. Wiesel, "Shape and arrangement of columns in cat's striate cortex," *The Journal of physiology*, vol. 165, no. 3, pp. 559–568, 1963.
- [87] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [88] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.
- [89] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.
- [90] M. C. Tjepkema-Cloostermans, R. C. de Carvalho, and M. J. van Putten, "Deep learning for detection of focal epileptiform discharges from scalp EEG recordings," *Clinical Neurophysiology*, 2018.
- [91] R. Xu, C. Li, A. H. Paterson, Y. Jiang, S. Sun, and J. S. Robertson, "Aerial images and convolutional neural network for cotton bloom detection," *Frontiers in Plant Science*, vol. 8, p. 2235, 2018. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fpls.2017.02235>
- [92] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [93] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.
- [94] M. P. Pound, J. A. Atkinson, A. J. Townsend, M. H. Wilson, M. Griffiths, A. S. Jackson, A. Bulat, G. Tzimiropoulos, D. M. Wells, E. H. Murchie *et al.*, "Deep machine learning provides state-of-the-art performance in image-based plant phenotyping," *GigaScience*, 2017.

- [95] E. D. Übeyli, “Recurrent neural networks employing lyapunov exponents for analysis of ecg signals,” *Expert Systems with Applications*, vol. 37, no. 2, pp. 1192–1199, 2010.
- [96] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen, “Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans,” *Scientific reports*, vol. 6, p. 24454, 2016.
- [97] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, p. 115, 2017.
- [98] J. A. Golden, “Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping artificial intelligence be seen,” *Jama*, vol. 318, no. 22, pp. 2184–2186, 2017.
- [99] T. X. Le, T. T. Le, V. V. Dinh, Q. L. Tran, L. T. Nguyen, and D. T. Nguyen, “Deep learning for epileptic spike detection,” *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 33, no. 2, pp. 1–13, 2018.
- [100] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [101] Y. Tang, “Deep learning using linear support vector machines,” *arXiv preprint arXiv:1306.0239*, 2013.
- [102] G. E. Hinton, “A practical guide to training restricted boltzmann machines,” in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 599–619.
- [103] N. T. Anh-Dao, N. Linh-Trung, L. Van Nguyen, T. Tran-Duc, B. Boashash *et al.*, “A multistage system for automatic detection of epileptic spikes,” *REV Journal on Electronics and Communications*, 2018.

- [104] P. Baldi, “Autoencoders, unsupervised learning, and deep architectures,” in *Proceedings of ICML workshop on unsupervised and transfer learning*, 2012, pp. 37–49.
- [105] A. Hyvärinen, “Survey on independent component analysis. neural computing surveys 2, 94–128 (1999).”
- [106] A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.
- [107] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
- [108] A. Hyvärinen, “Fast and robust fixed-point algorithms for independent component analysis,” *Neural Networks, IEEE Transactions on*, vol. 10, no. 3, pp. 626–634, 1999.
- [109] A. Hyvärinen and E. Oja, “A fast fixed-point algorithm for independent component analysis,” *Neural computation*, vol. 9, no. 7, pp. 1483–1492, 1997.
- [110] D. Pahm, P. Garrat, and C. Jutten, “Separation of a mixture of independent sources through a ml approach,” in *Proc. European Signal Processing Conf*, 1992, p. 771.
- [111] T.-P. Jung, S. Makeig, M. J. McKeown, A. J. Bell, T.-W. Lee, and T. J. Sejnowski, “Imaging brain dynamics using independent component analysis,” *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1107–1122, 2001.
- [112] G. Barbati, C. Porcaro, F. Zappasodi, P. M. Rossini, and F. Tecchio, “Optimization of an independent component analysis approach for artifact identification and removal in magnetoencephalographic signals,” *Clinical Neurophysiology*, vol. 115, no. 5, pp. 1220–1232, 2004.
- [113] M. L. Shoker, S. Sanei, W. Wang, and J. A. Chambers, “Removal of eye blinking artifact from the electro-encephalogram, incorporating a new constrained blind

- source separation algorithm,” *Medical and Biological Engineering and Computing*, vol. 43, no. 2, pp. 290–295, 2005.
- [114] R. Vigário, V. Jousmäki, M. Hämmäläinen, R. Haft, and E. Oja, “Independent component analysis for identification of artifacts in magnetoencephalographic recordings,” *Advances in neural information processing systems*, pp. 229–235, 1998.
 - [115] M. De Lucia, J. Fritschy, P. Dayan, and D. Holder, “The classification of spikes in EEG recordings using features derived from ica,” in *Advances in Medical, Signal and Information Processing, 2006. MEDSIP 2006. IET 3rd International Conference On*. IET, 2006, pp. 1–4.
 - [116] A. J. Bell and T. J. Sejnowski, “Learning the higher-order structure of a natural sound,” *Network: Computation in Neural Systems*, vol. 7, no. 2, pp. 261–266, 1996.
 - [117] M. I. Schlesinger and V. Hlavác, *Ten lectures on statistical and structural pattern recognition*. Springer Science & Business Media, 2013, vol. 24.
 - [118] P. Van Hese, B. Vanrumste, H. Hallez, G. J. Carroll, K. Vonck, R. D. Jones, P. J. Bones, Y. D’Asseler, and I. Lemahieu, “Detection of focal epileptiform events in the EEG by spatio-temporal dipole clustering,” *Clinical Neurophysiology*, vol. 119, no. 8, pp. 1756–1770, 2008.
 - [119] J. Hara, T. Musha, and W. R. Shankle, “Approximating dipoles from human EEG activity: the effect of dipole source configuration on dipolarity using single dipole models,” *Biomedical Engineering, IEEE Transactions on*, vol. 46, no. 2, pp. 125–129, 1999.
 - [120] D. Flanagan, R. Agarwal, Y. Wang, and J. Gotman, “Improvement in the performance of automated spike detection using dipole source features for artefact rejection,” *Clinical Neurophysiology*, vol. 114, no. 1, pp. 38–49, 2003.
 - [121] J. Gotman, P. Gloor, and N. Schaul, “Comparison of traditional reading of the EEG and automatic recognition of interictal epileptic activity,” *Electroencephalography and clinical Neurophysiology*, vol. 44, no. 1, pp. 48–60, 1978.

- [122] F. Sartoretto and M. Ermani, “Automatic detection of epileptiform activity by single-level wavelet analysis,” *Clinical Neurophysiology*, vol. 110, no. 2, pp. 239–249, 1999.
- [123] B. Ramabhadran, J. D. Frost Jr, J. R. Glover, and P. Y. Ktonas, “An automated system for epileptogenic focus localization in the electroencephalogram,” *Journal of clinical neurophysiology*, vol. 16, no. 1, pp. 59–68, 1999.
- [124] H. Goelz, R. D. Jones, and P. J. Bones, “Wavelet analysis of transient biomedical signals and its application to detection of epileptiform activity in the EEG,” *Clinical electroencephalography*, vol. 31, no. 4, pp. 181–191, 2000.
- [125] Z. H. İnan and M. Kuntalp, “A study on fuzzy c-means clustering-based systems in automatic spike detection,” *Computers in Biology and Medicine*, vol. 37, no. 8, pp. 1160–1166, 2007.
- [126] A. Nonclercq, M. Foulon, D. Verheulpen, C. D. Cock, M. Buzatu, P. Mathys, and P. V. Bogaert, “Cluster-based spike detection algorithm adapts to interpatient and inpatient variation in spike morphology,” *Journal of Neuroscience Methods*, vol. 210, no. 2, pp. 259 – 265, 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0165027012002798>
- [127] J. Zhang, J. Zou, M. Wang, L. Chen, C. Wang, and G. Wang, “Automatic detection of interictal epileptiform discharges based on time-series sequence merging method,” *Neurocomputing*, vol. 110, pp. 35 – 43, 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0925231212009150>
- [128] V. Chavakula, I. S. Fernández, J. M. Peters, G. Popli, W. Bosl, S. Rakhade, A. Rotenberg, and T. Loddenkemper, “Automated quantification of spikes,” *Epilepsy & Behavior*, vol. 26, no. 2, pp. 143–152, 2013.
- [129] S. S. Lodder and M. J. van Putten, “A self-adapting system for the automated detection of inter-ictal epileptiform discharges,” *PloS one*, vol. 9, no. 1, p. e85180, 2014.

- [130] R. Janca, P. Jezdik, R. Cmejla, M. Tomasek, G. A. Worrell, M. Stead, J. Wagenaar, J. G. Jefferys, P. Krsek, V. Komarek *et al.*, “Detection of interictal epileptiform discharges using signal envelope distribution modelling: application to epileptic and non-epileptic intracranial recordings,” *Brain topography*, vol. 28, no. 1, pp. 172–183, 2015.
- [131] K. Parthiban and S. Vijayachitra, “Spike detection from electroencephalogram signals with aid of hybrid genetic algorithm-particle swarm optimization,” *Journal of Medical Imaging and Health Informatics*, vol. 5, no. 5, pp. 936–944, 2015.
- [132] T. Fukami, T. Shimada, and B. Ishikawa, “Fast EEG spike detection via eigenvalue analysis and clustering of spatial amplitude distribution,” *Journal of neural engineering*, vol. 15, no. 3, p. 036030, 2018.
- [133] C. C. Pang, A. R. Upton, G. Shine, and M. V. Kamath, “A comparison of algorithms for detection of spikes in the electroencephalogram,” *Biomedical Engineering, IEEE Transactions on*, vol. 50, no. 4, pp. 521–526, 2003.
- [134] S. Wilson, R. Harner, F. Duffy, B. Tharp, M. Nuwer, and M. Sperling, “Spike detection. i. correlation and reliability of human experts,” *Electroencephalography and clinical Neurophysiology*, vol. 98, no. 3, pp. 186–198, 1996.
- [135] A. J. Casson, E. Luna, and E. Rodriguez-Villegas, “Performance metrics for the accurate characterisation of interictal spike detection algorithms,” *Journal of neuroscience methods*, vol. 177, no. 2, pp. 479–487, 2009.
- [136] W. E. Hostetler, H. J. Doller, and R. W. Homan, “Assessment of a computer program to detect epileptiform spikes,” *Electroencephalography and clinical neurophysiology*, vol. 83, no. 1, pp. 1–11, 1992.
- [137] M. L. Scheuer, A. Bagic, and S. B. Wilson, “Spike detection: Inter-reader agreement and a statistical turing test on a large data set,” *Clinical Neurophysiology*, vol. 128, no. 1, pp. 243–250, 2017.
- [138] J. Jing, J. Dauwels, T. Rakthanmanon, E. Keogh, S. Cash, and M. Westover, “Rapid annotation of interictal epileptiform discharges via template matching un-

- der dynamic time warping,” *Journal of neuroscience methods*, vol. 274, pp. 179–190, 2016.
- [139] R. Esteller, J. Echauz, T. Tchong, B. Litt, and B. Pless, “Line length: an efficient feature for seizure onset detection,” in *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, vol. 2. IEEE, 2001, pp. 1707–1710.
- [140] P. Maragos, J. F. Kaiser, and T. F. Quatieri, “On amplitude and frequency demodulation using energy operators,” *IEEE Transactions on signal processing*, vol. 41, no. 4, pp. 1532–1550, 1993.
- [141] J. F. Kaiser, “On a simple algorithm to calculate the ‘energy’ of a signal,” in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990, pp. 381–384.
- [142] S. Mukhopadhyay and G. Ray, “A new interpretation of nonlinear energy operator and its efficacy in spike detection,” *Biomedical Engineering, IEEE Transactions on*, vol. 45, no. 2, pp. 180–187, 1998.
- [143] J. Choi and T. Kim, “Neural action potential detector using multi-resolution teo,” *Electronics Letters*, vol. 38, no. 12, p. 1, 2002.
- [144] A. Mertins and D. A. Mertins, *Signal analysis: wavelets, filter banks, time-frequency transforms and applications*. John Wiley & Sons, Inc., 1999.
- [145] J. D. Kelleher, B. Mac Namee, and A. D’Arcy, *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press, 2015.
- [146] E. T. Berkman and S. P. Reise, *A conceptual guide to statistics using SPSS*. Sage, 2011.
- [147] R. L. Wasserstein and N. A. Lazar, “The asa’s statement on p-values: Context, process, and purpose,” *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016. [Online]. Available: <https://doi.org/10.1080/00031305.2016.1154108>

- [148] W. Webber and R. P. Lesser, "Automated spike detection in EEG," *Clinical Neurophysiology*, vol. 128, no. 1, pp. 241–242, 2017.
- [149] J. Jin, J. Dauwels, S. Cash, and M. B. Westover, "Spikegui: Software for rapid interictal discharge annotation via template matching and online machine learning," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 4435–4438.
- [150] J. J. Halford, A. Arain, G. P. Kalamangalam, S. M. LaRoche, B. Leonardo, M. Basha, N. J. Azar, E. Kutluay, G. U. Martz, W. J. Bethany *et al.*, "Characteristics of EEG interpreters associated with higher interrater agreement," *Journal of Clinical Neurophysiology*, vol. 34, no. 2, pp. 168–173, 2017.
- [151] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.
- [152] J. Uebersax, "Kappa coefficients: A critical appraisal," *Statistical methods for rater agreement*, 2002.
- [153] E. Blood and K. F. Spratt, "Disagreement on agreement: Two alternative agreement coefficients," in *SAS Global Forum*. Citeseer, 2007.
- [154] J. Uebersax, "Measuring diagnostic reliability.[reply to spitznagel and helzer letter]," *Arch Gen Psychiatry*, vol. 44, pp. 193–4, 1987.
- [155] A. R. Feinstein and D. V. Cicchetti, "High agreement but low kappa: I. the problems of two paradoxes," *Journal of clinical epidemiology*, vol. 43, no. 6, pp. 543–549, 1990.
- [156] W. D. Thompson and S. D. Walter, "A reappraisal of the kappa coefficient," *Journal of clinical epidemiology*, vol. 41, no. 10, pp. 949–958, 1988.
- [157] —, "Response kappa and the concept of independent errors," *Journal of clinical epidemiology*, vol. 41, no. 10, pp. 969–970, 1988.

- [158] K. Gwet, “Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters,” *Gaithersburg, MD: STATAXIS Publishing Company*, 2001.
- [159] A. Stuart, “A test for homogeneity of the marginal distributions in a two-way classification,” *Biometrika*, vol. 42, no. 3/4, pp. 412–416, 1955.
- [160] A. E. Maxwell, “Comparing the classification of subjects by two independent judges,” *The British Journal of Psychiatry*, vol. 116, no. 535, pp. 651–655, 1970.
- [161] K. L. Gwet, “Computing inter-rater reliability and its variance in the presence of high agreement,” *British Journal of Mathematical and Statistical Psychology*, vol. 61, no. 1, pp. 29–48, 2008.
- [162] —, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.
- [163] N. Wongpakaran, T. Wongpakaran, and K. L. Gwet, “A comparison of cohen’s kappa and gwet’s ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples,” *BMC medical research methodology*, vol. 13, no. 1, p. 61, 2013.
- [164] M. R. Nuwer, D. Lehmann, F. L. da Silva, S. Matsuoka, W. Sutherling, and J.-F. Vibert, “Ifcn guidelines for topographic and frequency analysis of EEGs and eps. report of an ifcn committee,” *Electroencephalography and clinical Neurophysiology*, vol. 91, no. 1, pp. 1–5, 1994.
- [165] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.
- [166] N. Kingsbury, “Complex wavelets for shift invariant analysis and filtering of signals,” *Applied and computational harmonic analysis*, vol. 10, no. 3, pp. 234–253, 2001.
- [167] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, “The dual-tree complex wavelet transform,” *IEEE signal processing magazine*, vol. 22, no. 6, pp. 123–151, 2005.

- [168] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," in *IEEE Trans. on Acoust., Speech, and Sig. Proc.*, vol. 4, 1997, pp. 21–24.
- [169] F. Éva, "Statistical analysis of the factors which take effect on the new huf credit volume in the sme sector," *Proceedings in ARSA-Advanced Research in Scientific Areas*, no. 1, 2012.
- [170] C. F. Dietrich, *Uncertainty, calibration and probability: the statistics of scientific and industrial measurement*. CRC Press, 1991.
- [171] C. M. Abraham and M. S. Elayidom, "Implementation of correlation techniques for the equity market in india."
- [172] E. C. Fieller, H. O. Hartley, and E. S. Pearson, "Tests for rank correlation coefficients. i," *Biometrika*, vol. 44, no. 3/4, pp. 470–481, 1957.
- [173] A. Lehman, *JMP for basic univariate and multivariate statistics: a step-by-step guide*. SAS Institute, 2005.
- [174] J. L. Myers, A. Well, and R. F. Lorch, *Research design and statistical analysis*. Routledge, 2010.
- [175] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.
- [176] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [177] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
- [178] P. Schiilkop, C. Burgest, and V. Vapnik, "Extracting support data for a given task," in *Proceedings of the 1st international conference on knowledge discovery & data mining*, 1995, pp. 252–257.

- [179] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, “New support vector algorithms,” *Neural computation*, vol. 12, no. 5, pp. 1207–1245, 2000.
- [180] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [181] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.
- [182] S. K. M. Yi, M. Steyvers, M. D. Lee, and M. J. Dry, “The wisdom of the crowd in combinatorial problems,” *Cognitive science*, vol. 36, no. 3, pp. 452–470, 2012.
- [183] B. Santoshkumar, J. J. Chong, W. T. Blume, R. S. McLachlan, G. B. Young, D. C. Diosy, J. G. Burneo, and S. M. Mirsattari, “Prevalence of benign epileptiform variants,” *Clinical Neurophysiology*, vol. 120, no. 5, pp. 856–861, 2009.
- [184] P. D. Corporation. (2016, Dec.) Persyst 13 has broken the code on spike detection. [Online]. Available: <https://www.persyst.com/technology/spike-detection/>
- [185] M. B. Westover, J. J. Halford, and M. T. Bianchi, “What it should mean for an algorithm to pass a statistical turing test for detection of epileptiform discharges,” *Clinical Neurophysiology*, vol. 128, no. 7, pp. 1406–1407, 2017.
- [186] K. Kelly, D. Shiau, R. Kern, J. Chien, M. Yang, K. Yandora, J. Valeriano, J. Halford, and J. Sackellares, “Assessment of a scalp EEG-based automated seizure detection system,” *Clinical Neurophysiology*, vol. 121, no. 11, pp. 1832–1843, 2010.
- [187] J. J. Halford, M. R. Sperling, D. R. Nair, D. J. Dlugos, W. O. Tatum, J. Harvey, J. A. French, J. R. Pollard, E. Faught, K. H. Noe *et al.*, “Detection of generalized tonic–clonic seizures using surface electromyographic monitoring,” *Epilepsia*, vol. 58, no. 11, pp. 1861–1869, 2017.
- [188] J. J. Halford, M. B. Westover, S. M. LaRoche, M. P. Macken, E. Kutluay, J. C. Edwards, L. Bonilha, G. P. Kalamangalam, K. Ding, J. L. Hopp *et al.*, “Interictal epileptiform discharge detection in EEG in different practice settings.” *Journal of*

clinical neurophysiology: official publication of the American Electroencephalographic Society, 2018.

- [189] V. Shankar and S. I. Bangdiwala, “Observer agreement paradoxes in 2x2 tables: comparison of agreement measures,” *BMC medical research methodology*, vol. 14, no. 1, p. 100, 2014.
- [190] S. Wellek, *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall/CRC, 2010.
- [191] W. O. Tatum, “How not to read an EEG introductory statements,” *Neurology*, vol. 80, no. 1 Supplement 1, pp. S1–S3, 2013.