




2020

TEMPORAL DATA EXTRACTION AND QUERY SYSTEM FOR EPILEPSY SIGNAL ANALYSIS

Yan Huang

University of Kentucky, yhu247@uky.edu

Author ORCID Identifier:

 <https://orcid.org/0000-0002-6578-2379>

Digital Object Identifier: <https://doi.org/10.13023/etd.2020.182>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Huang, Yan, "TEMPORAL DATA EXTRACTION AND QUERY SYSTEM FOR EPILEPSY SIGNAL ANALYSIS" (2020). *Theses and Dissertations--Computer Science*. 98.
https://uknowledge.uky.edu/cs_etds/98

This Doctoral Dissertation is brought to you for free and open access by the Computer Science at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Computer Science by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@sv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Yan Huang, Student

Dr. Guo-Qiang Zhang, Major Professor

Dr. Miroslaw Truszczyński, Director of Graduate Studies

TEMPORAL DATA EXTRACTION AND QUERY SYSTEM FOR EPILEPSY
SIGNAL ANALYSIS

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Engineering
at the University of Kentucky

By

Yan Huang

Lexington, Kentucky

Co-Directors: Dr. Guo-Qiang Zhang, Professor of Computer Science
and Dr. Jin Chen, Associate Professor of Computer Science

Lexington, Kentucky

Copyright © Yan Huang 2020

<https://orcid.org/0000-0002-6578-2379>

ABSTRACT OF DISSERTATION

TEMPORAL DATA EXTRACTION AND QUERY SYSTEM FOR EPILEPSY SIGNAL ANALYSIS

The 2016 Epilepsy Innovation Institute (Ei²) community survey reported that unpredictability is the most challenging aspect of seizure management. Effective and precise detection, prediction, and localization of epileptic seizures is a fundamental computational challenge. Utilizing epilepsy data from multiple epilepsy monitoring units can enhance the quantity and diversity of datasets, which can lead to more robust epilepsy data analysis tools. The contributions of this dissertation are two-fold. One is the implementation of a temporal query for epilepsy data; the other is the machine learning approach for seizure detection, seizure prediction and seizure localization. The three key components of our temporal query interface are: 1) A pipeline for automatically extract European Data Format (EDF) information and epilepsy annotation data from cross-site sources; 2) Data quantity monitoring for Epilepsy temporal data; 3) A web-based annotation query interface for preliminary research and building customized epilepsy datasets. The system extracted and stored about 450,000 epilepsy-related events of more than 2,497 subjects from seven institutes up to September 2019. Leveraging the epilepsy temporal events query system, we developed machine learning models for seizure detection, prediction, and localization. Using 135 extracted features from EEG signals, we trained a channel-based eXtreme Gradient Boosting model to detect seizures on 8-second EEG segments. A long-term EEG recording evaluation shows that the model can detect about 90.34% seizures on an existing EEG dataset with 961 hours of data. The model achieved 89.88% accuracy, 92.32% sensitivity and 84.76% AUC based on the segments evaluation. We also introduced a transfer learning approach consisting of 1) a base deep learning model pre-trained by ImageNet dataset and 2) customized fully connected layers, to train the patient-specific pre-ictal and inter-ictal data from our database. Two convolutional neural network architectures were evaluated using 53 pre-ictal segments and 265 continuous hours of inter-ictal EEG data. The evaluation shows that our model reached 86.79% sensitivity and 3.38% false-positive rate. Another transfer learning model for seizure localization uses a pre-trained ResNext50 structure and was trained with an image augmentation dataset labeling by fingerprint. Our model achieved

88.22% accuracy, 34.99% sensitivity, 1.02% false-positive rate, and 34.3% positive likelihood rate.

KEYWORDS: EEG, Temporal Query, Data Quality, Seizure Detection, Seizure Prediction, Seizure Localization

YAN HUANG

Student's Signature

APRIL 28, 2020

Date

TEMPORAL DATA EXTRACTION AND QUERY SYSTEM FOR EPILEPSY
SIGNAL ANALYSIS

By
Yan Huang

GUO-QIANG ZHANG

Co-Director of Dissertation

JIN CHEN

Co-Director of Dissertation

MIROSLAW TRUSZCZYNSKI

Director of Graduate Studies

APRIL 28, 2020

Date

Dedicated to my beloved parents and my wife.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor Prof. GQ Zhang for bringing me to the world of biomedical informatics and his consistent support for my graduate study and life. Prof. Zhang always convincingly guides me to be professional and proactive even at the difficult time of projects development or paper writing. He provides a comfortable scientific research environment for the group and encourages me to open my mind to new research topics. Without his persistent mentorship, my works and the dissertation would not have been reached the goals.

I would also like to thank the rest of my committee: chair Prof. Jin Chen, co-chair Prof. Jinze Liu, committee member Prof. Tingting Yu and Prof. David Fardo. Thank you for been my dissertation readers and providing insightful comments and inspirations in many ways. I am also thankful for all the professors of the courses I took at Case Western Reserve University and the University of Kentucky. Their unsurpassed teaching helped me a lot during my research.

To the organizations: National Sleep Research Resource and the Center for SUDEP Research, without their resources and funding, this dissertation could not have been accomplished. To the domain experts from Institute for Biomedical Informatics at University of Kentucky and the University of Texas Health Science Center at Houston: their valuable domain knowledge and clinical experience is really important to my projects. I would also like to express my special thanks to Dr. Samden Lhatoo, Dr. Stephen Thompson, Dr. John Mosher, and Dr. Shirin Omid, their inputs and feedbacks help me to improve the development of my tools.

Besides, I would also like to acknowledge all the members in our group. Thanks to Dr. Licong Cui and Dr. Shiqiang Tao for the discussions on every problem, experiment and solution of my projects. Thanks to my friends Dr. Wei Zhu, Dr.

Xiaojin Li, Dr. Ningzhou Zeng and Xi Wu for all the hard working and fun time together.

Last but not the least, I feel a deep sense of grateful for my family, nobody has been more important to me than them. Thanks to my mother and father for giving birth to me and teaching me all the important things that matter in my life. To my beloved wife, Shan, your encouragement and support are always the strength of me.

Table of Contents

Acknowledgements	iii
List of Tables	viii
List of Figures	x
1 Introduction	1
1.1 Epilepsy Temporal Data	3
1.2 EEG Annotation Visualization Tools	7
1.3 Machine Learning on EEG Signals	8
1.4 Contribution	10
1.5 Outline	12
2 Background	14
2.1 Epilepsy Data Sources	14
2.1.1 Public Epilepsy Datasets	14
2.1.2 Center for SUDEP Research	15
2.2 Cross-site Epilepsy Data Management	16
2.2.1 Epilepsy Ontology	19
2.2.2 FAIR Data Principle	19
2.2.3 Cross-site Epilepsy Data Capture and Integration	20
2.2.4 Data Quality Assurance	21
2.2.5 EEG Information Retrieval	23
2.3 Machine Learning Methods	23
2.3.1 eXtreme Gradient Boosting (XGBoost)	23
2.3.2 Convolutional Neural Network (CNN)	24
2.3.3 Transfer Learning	25
2.4 EEG Signal Classification	25
2.4.1 EEG Signal Data Processing	26
2.4.2 Seizure Detection	27
2.4.3 Seizure Prediction	27
2.4.4 Epileptogenic Zone in Epilepsy	28
2.5 Evaluation of EEG Signal Classification	29
2.5.1 Evaluation Data	30
2.5.2 Evaluation Measurements	31
2.6 Development Environments	33
2.6.1 Ruby on Rails	33
2.6.2 MySQL	34
2.6.3 TensorFlow	35

3	Temporal Query for Epilepsy Dataset	36
3.1	Motivation	36
3.2	Dataset	39
3.3	EEG Temporal Data Extraction	41
3.4	Temporal Query	43
3.5	Results	49
3.5.1	CSR Temporal Data Quality	49
3.5.2	Temporal Query User Interface Design	51
3.5.3	Customized Epilepsy Dataset Builder	56
3.6	Discussion	58
3.7	Conclusion	59
4	Seizure Detection on Scalp EEG Data	60
4.1	Motivation	60
4.2	Datasets	62
4.3	Pre-processing	65
4.4	EEG Signal Classification	69
4.5	Result	71
4.6	Discussion	73
4.7	Conclusion	74
5	Seizure Prediction on Scalp EEG Data	76
5.1	Motivation	76
5.2	Datasets	78
5.3	Pre-processing	81
5.4	Classification Model	82
5.5	Result	86
5.6	Discussion	87
5.7	Conclusion	89
6	Seizure Localization on Stereoelectroencephalography Data	90
6.1	Motivation	90
6.2	Datasets	93
6.3	Pre-processing	94
6.3.1	Channel-based Segmentation	94
6.3.2	Time-frequency Image Generation	96
6.3.3	Image Augmentation	97
6.4	Classification Model	98
6.5	Result	102
6.6	Discussion	105
6.7	Conclusion	106

7	Conclusions and Future Work	108
7.1	Conclusions	108
7.2	Future Work	111
7.2.1	A More Powerful Cloud-based EEG interface	111
7.2.2	EEG Signal Quality Assurance	111
7.2.3	Features Enhancement on EEG Signal Analysis	112
	REFERENCES	113
	Vita	120

List of Tables

2.1	A comparison between CSR and other public datasets: University Hospital Bonn Germany seizure dataset, Children’s Hospital Boston-Massachusetts Institute of Technology, and 2014 Kaggle seizure prediction contest dataset. *Only includes the subjects with annotated seizures.	16
2.2	Data completeness for ten common data elements in patient reports from multiple sites.	22
2.3	The data collection method of 9 published seizure detection methods.	31
2.4	The evaluation measurements for the same model on different evaluation sets.	32
3.1	Retrieve information from University of Bonn EEG dataset using labels. The four characteristics are not in common for all labels and they are used to distinguish between each other.	38
3.2	EDF header structure	40
3.3	EEG signal files completeness for patients and patient reports from multiple sites	50
3.4	Data quality measurement for EEG signal files from multiple sites . .	50
3.5	Data quality measurement for EEG annotations from multiple sites .	51
3.6	Long-term EEG monitoring data completeness and duplication	51
3.7	The completeness comparison between the CHB-MIT scalp EEG dataset and a 23-patient subset of the CSR dataset.	57
4.1	The details of collected data of 23 patients from the CHB-MIT dataset and 23 patients from the CSR dataset. SZ = number of seizures, LSZ = number of lead seizures.	64
4.2	CHB-MIT dataset testing results.	68
4.3	CSR dataset testing results.	72
4.4	Comparison table for testing results on CHB-MIT dataset.	74
5.1	The summary of the 17 patients randomly selected from the CSR dataset.	80
5.2	The performance of the transfer learning with two base models: VGG19, ResNeXt50. Two measurements are Sensitivity and FPR (False Positive Rate).	85
6.1	The details of collected data of five patients.	94
6.2	Model performance comparison. cnn : convolutional neural network model. resnext50 : resnext50 transfer learning model. na : unbalanced non-augmentation dataset. a : balanced augmentation dataset. fp: labeling using fingerprint wavelet pattern. r : labeling using resection zone	102
6.3	Performance comparison between EZ-Fingerprint and this work. . .	103
6.4	Case study on Patient-1 and Patient-5.	106

6.5 Process running time of the two pipelines. The testing machine is MacBook Pro 2015 with 2.7GHz Intel Core i5 CPU, and 8GB RAM. 106

List of Figures

1.1	Electrode labeling in International 10-20 system with a 18-Channel longitudinal bipolar montage.	3
1.2	Examples of a 30 seconds EEG signal clip of channel F3 in three categories : (a) Inter-ictal, (b) Pre-ictal, and (c) Ictal. The waves are visualized by SeizureBank [5].	4
1.3	Epilepsy data explosion in CSR from February 2016 to September 2019.	5
1.4	Use the NSRR EDF Viewer to visualize signals and annotation by selecting a demand annotation in the right hand side box.	7
1.5	A concepts overview of this dissertation. TeQ: Temporal events Query. EpiD: Epilepsy Detection. EpiP: Epilepsy Prediction. EpiL: Epileptogenic zone Localization.	11
2.1	10-hour continuous evaluations of seizure detection model A and seizure detection model B. Both of the models have three false alarms and one correct detection.	33
3.1	The general architecture of an ontology-driven information extraction system.	42
3.2	Ontology-driven annotation information extraction for CSR.	43
3.3	Ontology vocabulary for EEG annotation.	44
3.4	Expression tree representation for: “Sign of Four” is during “Clinical Seizure Onset” and “Clinical Seizure End”.	45
3.5	Example of a historical expression tree.	46
3.6	CSR temporal annotation query database design.	48
3.7	CSR Temporal annotation query interface.	52
3.8	An example of zoom out function.	54
3.9	An example of zoom in function.	54
3.10	An example of showing all standard annotations.	55
3.11	An example of showing all annotations.	55
5.1	An example of 15-channel spectrogram images. For each image, the x-axis is frequency and the y-axis is time. A red point indicates higher energy at the time and frequency, and the blue means a lower energy point. At around 60Hz, a power line exists in most images. Such noise is eliminated during data pre-processing.	77
5.2	An example of our pre-ictal segments and inter-ictal segments extraction method. The horizontal black line is the timeline. The vertical red lines indicate the start points of seizure. The red dashed boxes cover the periods of the pre-ictal segments, and the green dashed box covers the period of the inter-ictal segment.	79

5.3	An example of prediction decision queue with threshold of 8. At the top is the time-series prediction results of input clips using pre-ictal detection model. The dotted box displays the status of the prediction decision queue at t_1, t_2 and t_3 , and the solid line boxes on the left show the final prediction results at t_1, t_2 and t_3	83
5.4	Monitoring of the evaluation results of patient 2 using ResNeXt50 transfer learning model : (a) a 1-hour pre-ictal testing segment, (b) A 2.5-hour inter-ictal testing segment.	87
6.1	Time domain SEEG signals of two channels in a 40-second window. The channel in green is outside the epileptogenic zone and the channel in red is outside the epileptogenic zone. The red dotted line in the middle denotes the start point of the seizure.	91
6.2	Time-frequency domain representation for SEEG signals in a 40-seconds window. The x-axis is the time axis ranges from 0 - 40 seconds and the y-axis is the frequency axis ranges from 0 - 200Hz.	92
6.3	Two “bad” fingerprint examples.	93
6.4	Workflow of SEEG signal data pre-processing.	95
6.5	Channel-based segmentation on channel RSMA2-RSMA3 from Patient-1.	96
6.6	Examples of five augmentation methods performed on a original wavelet image. The image is created by signals from channel RSMA2-RSMA3 of Patient-1.	99
6.7	Stacked CNN model structure.	100
6.8	Transfer learning model structure using ResNext50.	100
6.9	Venn diagram of overall performance for EZ-Fingerprint and this work.	104

CHAPTER 1. Introduction

Epilepsy is a central nervous system disorder that affects all genders and races and can occur at any age. An epilepsy patient is diagnosed by having multiple seizures on multiple occasions. According to the data from the Centers for Disease Control and Prevention (CDC), 1.2% of the US population had active epilepsy in 2015 [1]. Based on the current study, epilepsy is caused by an electrical disturbance in the brain, which may lead to uncontrollable behavior and loss of consciousness. The risk may become higher when the patients are performing daily routines, such as diving or swimming. More pressingly, the leading cause of epilepsy-related death, Sudden Unexpected Death in Epilepsy (SUDEP), is a life-threatening risk that may occur in patients with intractable, frequent, and continuing seizures [2]. Although most seizures naturally cease with no danger, a system warning within a meaningful lead time is an effective way to prevent harms. However, for many years, epileptic seizures have been marked as an “unpredictable” disorder because no available tools were available to reliably predict seizure onset.

A seizure is a sudden, uncontrolled abnormal brain activity which may lead to signs such as untypical whole or partial body movements. The gold standard of epilepsy diagnosis is using an electroencephalogram (EEG) to find special brain wave patterns. In the recent decades, many long-term brain activities of seizure patients have been captured by monitoring patients’ electrophysiological status. Two commonly used EEG types in the clinical settings are scalp EEG and stereoelectroencephalography (SEEG). EEG records voltage fluctuations resulting from ionic current within the neurons of the brain [3] while scalp EEG puts the electrodes on the surface of the brain and SEEG inserts the electrodes deep inside the brain. An EEG recording system typically consists of tens of EEG electrodes, and each electrode represents a continuing voltage signal at a specific brain location. The name of each electrode is defined by

the location according to a standard international method, the International 10-20 system [4], which uses the Latin alphabet to indicate the area of a brain: Pre-frontal (Fp), Frontal (F), Temporal (T), Parietal (P), Occipital (O), and Central (C). Even numbers refer to the right side of a brain and odd numbers refer to the left side of a brain (see Figure 1.1). Figure 1.2 illustrates the visualization of digital EEG signal data from channel F3 in three phases: ictal is the duration that a seizure occurs; pre-ictal is a period before seizure onset, and inter-ictal is the section other than ictal or pre-ictal.

By analyzing EEG signal waves, experts can identify abnormal patterns before or during seizures. Since visual inspection is ineffective, automatic algorithms have been developed for EEG signal classification including seizure detection, seizure prediction, and seizure localization. Because seizure is sudden and unpredictable, epilepsy care with long-term EEG monitoring is necessary for seizure control and treatment. As a result, many prospective EEG datasets especially seizure datasets have been collected in epilepsy monitoring center (EMU) across the states. With the rapid growth of “big data”, combining machine learning and data management is becoming a desirable solution for signal analysis on epilepsy data.

In this dissertation, we describe an end-to-end pipeline for a machine learning approach on EEG signal analysis. The major components of the temporal events query module include an ontology guided epilepsy temporal data extraction and integration system, and a web-based graphical user interface. We also demonstrate how to use the interface to create a high-quality EEG dataset with a specific task. Moreover, we introduce automatic algorithms for scalp EEG seizure detection, scalp EEG seizure prediction and SEEG epileptogenic zone localization. Evaluations for seizure detection and prediction on long-term EEG are reported.

To begin with, we describe different types of epilepsy temporal data and current challenges on extraction, management, and retrieval. We also review recent tools for

EEG annotation visualization and query. Besides, we summarize the related work on EEG signal classification using machine learning and potential benefit in the clinical aspect by using such tools. Finally, we highlight the contribution of this dissertation on the field of epilepsy EEG temporal data processing and analysis.

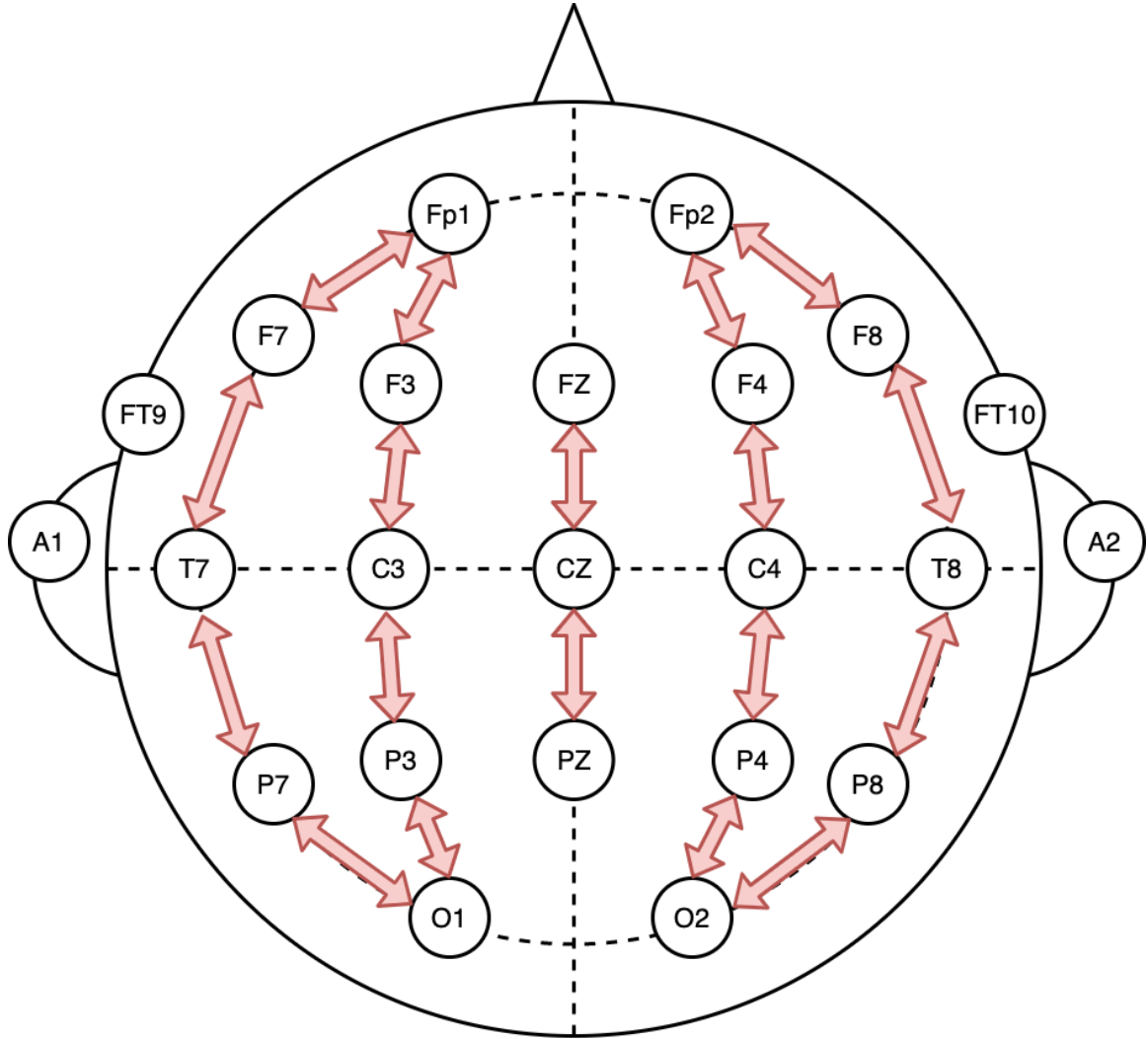


Figure 1.1: Electrode labeling in International 10-20 system with a 18-Channel longitudinal bipolar montage.

1.1 Epilepsy Temporal Data

After German physiologist and psychiatrist Hans Berger invented the electroencephalogram and recorded the first segment of the human EEG signal, EEG recording has

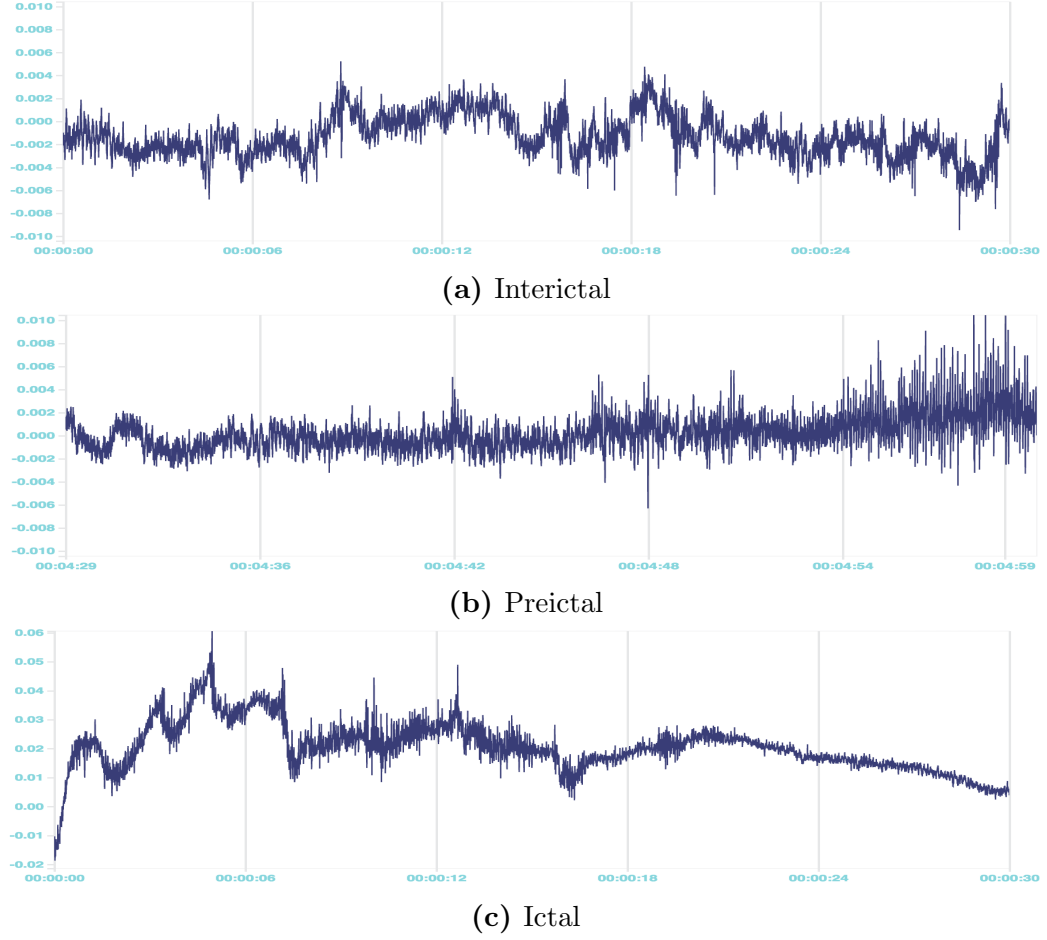


Figure 1.2: Examples of a 30 seconds EEG signal clip of channel F3 in three categories : (a) Inter-ictal, (b) Pre-ictal, and (c) Ictal. The waves are visualized by SeizureBank [5].

become one of the most convincing methods in the area of clinical neurology. The EEG signals are the most important temporal data in epilepsy study. Each point of an EEG recording has no obvious meaning but we can recognize the changes of frequency and amplitude with a time series of EEG signals. Some detailed abnormal wave patterns can be identified by experts and used for an epilepsy diagnosis. The EEG signals data belongs to one type of epilepsy temporal data, all three types are:

- Structured epilepsy temporal data which has a self-defined meaning and a fixed dimension for each data field, for example, data in the EDF header and report.
- Semi-structured epilepsy temporal data which has no fixed dimension but a well-

defined meaning for each data field, for example, annotation of the recordings.

- Unstructured epilepsy temporal data, which has neither fixed dimension nor well-defined meaning for each data field, for example, video and EEG monitoring.

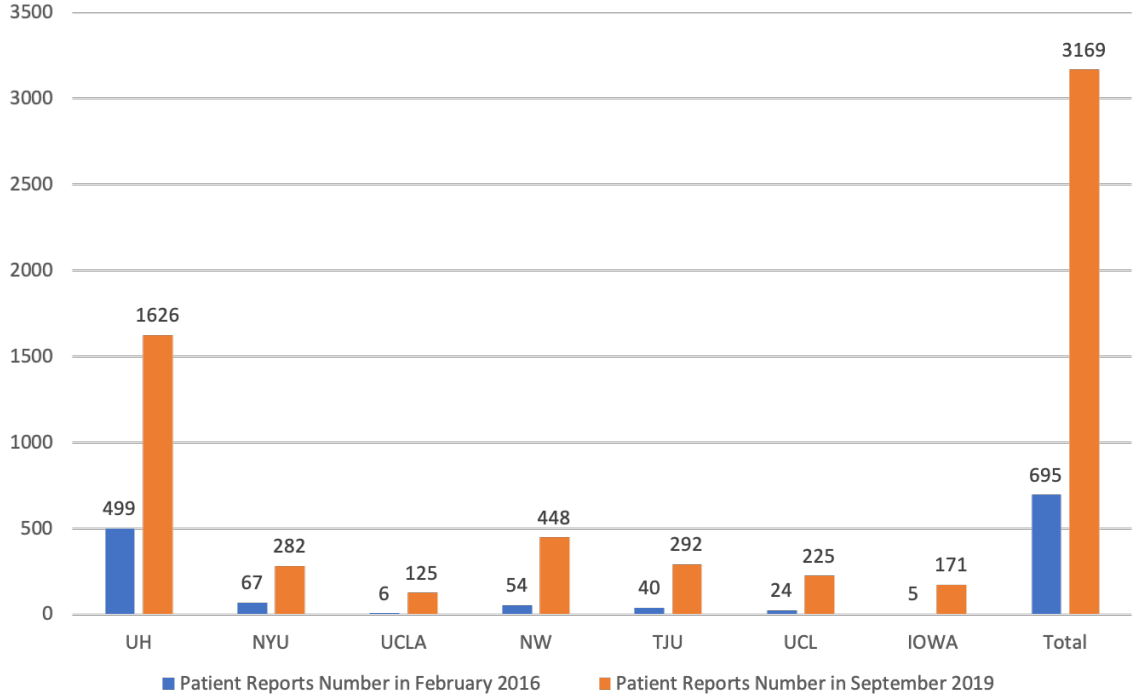


Figure 1.3: Epilepsy data explosion in CSR from February 2016 to September 2019.

An epilepsy temporal information system aims to extract and organize the information from the structured and semi-structured epilepsy temporal data for better describing the EEG signals. A well-described EEG signals dataset can be used for analyzing and understanding the seizure activities. However, in the case of seizure-related study, the probability of catching an EEG recording with seizure events is extremely low because of its short duration and rareness. Nowadays, EMU around the world is recording continuing EEG data that last for days or weeks. Center for SUDEP Research (CSR) is a collaboration of expertise from 7 institutions across the U.S. and Europe. A central database collects epilepsy data from collaborative clin-

ical sites prospectively. CSR dataset includes EEG signal data of more than 1,000 subjects, which is much larger than the existing public EEG seizure dataset in the number of patients. Figure 1.3 shows the data explosion in CSR from February 2016 to September 2019. The study contains 450,000 pieces of temporal annotation including machine codes and free text come with the EEG signal data. The experts are prospectively adding annotation by reviewing EEG signal patterns and patients' monitoring videos.

Combining the large scale epilepsy temporal data of EMUs from multiple institutions can enhance the data quantity and diversity, but how to control the data quality is another problem. Moreover, there are no existing temporal functions that can retrieve the epilepsy information across the subjects or recordings from different data source sites. To leverage such big EEG signal data into epilepsy research, several challenges must be overcome, which include but are not limited to:

- Data integration and management. The data format from different sources and sites may be different. Different experts may also use different terms to annotate the EEG data. A unified platform for users to upload, curate and explore across those epilepsy data is necessary for the large scale epilepsy information system.
- Multi-site information extraction. Epilepsy information is collected from three sources: EDF files (EEG signals and metadata), annotation text files, and subjects' demographic data from EHRs. To ensure the extracted information is updated, an automatic method is needed to perform recurrence extraction tasks. Also, the extraction has to consider how to combine and unify the data from different institutions.
- Time series information retrieval. EEG signals are time-series data and EEG annotations are timestamp data. One major purpose of Epilepsy information is to locate EEG signals segments using annotation information, which requires

an effective algorithm and a human-friendly query interface for complicated temporal queries.

1.2 EEG Annotation Visualization Tools

EEG annotation is a type of semi-structured temporal data embedded in EEG signals. An EEG annotation can be either a text description of a time point or a time interval. The visualization of EEG annotation in a timeline offers an intuitive understanding of temporal correlation of events during the recording. Existing EEG annotation visualization is usually an add-on feature of EEG signal visualization tools. The main purpose of the tool is retrieving the details of EEG signals, while loading or editing annotation is optional.

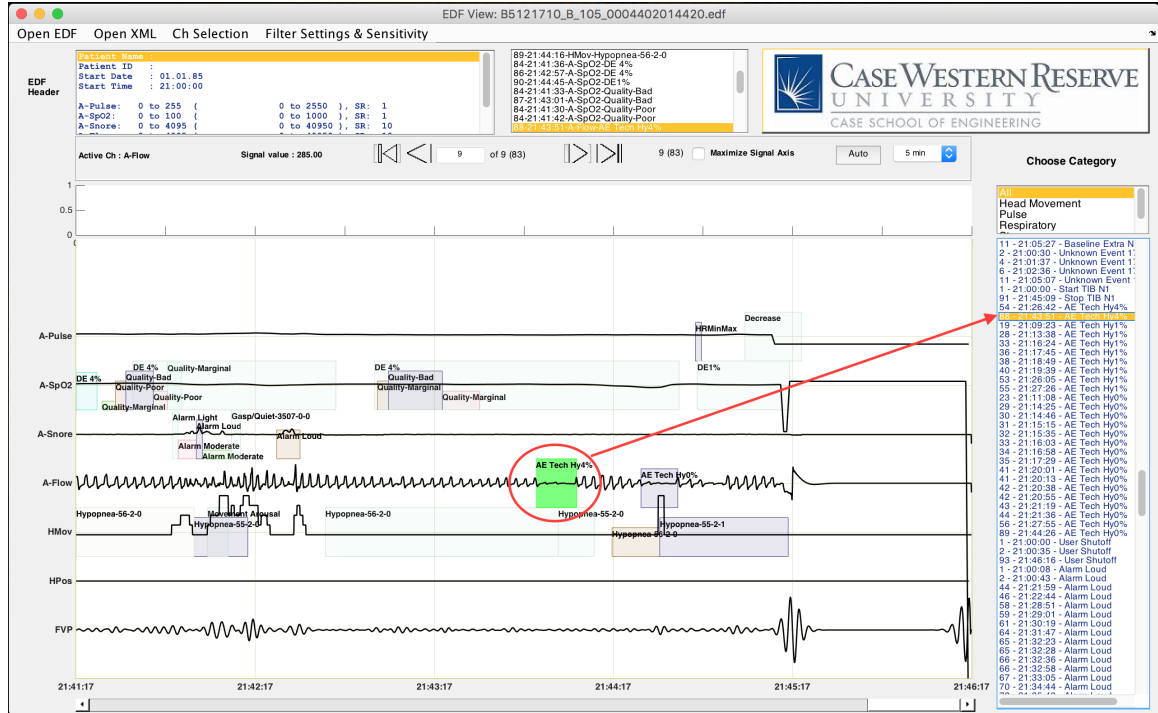


Figure 1.4: Use the NSRR EDF Viewer to visualize signals and annotation by selecting a demand annotation in the right hand side box.

National Sleep Research Resource (NSRR) Cross Cohort EDF Viewer (Figure 1.4) is an EEG visualization tool for EDF files. The EDF Viewer was developed by a

collaboration between Case Western Reserve University and Brigham and Women’s Hospital to extend the original Physio-MIMI viewer to an open-source MATLAB EDF viewer. The viewer enables the user to open an EDF file and a corresponding sleep annotation file, which contains sleep scoring information. The user can select the signals to view and a myriad of ways are provided including scrolling through the signals, clicking on the Hypnogram, and clicking on a specific annotation.

Other EEG visualization software, such as EEGLAB, SigViewer, and MNE (MEG+EEG Analysis and visualization) also have similar features as the NSRR EDF viewer. Their common limitations are:

- Other software or external dependencies are required in the local computer environment;
- EEG signal files need to be prepared in local machine, such files are usually large in storage; and
- No annotation search function is provided, users may struggle when the number of annotation is large.

1.3 Machine Learning on EEG Signals

With the large scale Epilepsy data, machine learning is a suitable approach to perform certain tasks such as building EEG signals classifiers. Three major EEG signals classification problems are seizure detection, seizure prediction, and seizure localization. Starting with seizure detection on EEG signals in the 1970s, researchers successfully extracted relevant seizure features to recognize a seizure from EEG recording [6, 7], but the early study on EEG-based seizure detection only had 22% accuracy [6]. Other than time-domain study, EEG signal data is also analyzed in the frequency domain for classification [8, 9]. In recent years, traditional machine learning methods, such as support vector machine (SVM) and random forest, made huge progress on seizure

classification and prediction [10, 11]. In 2018, perfect results (100% accuracy, 100% sensitivity and 100% specificity) were achieved by an inter-patient model [12]. Additionally, deep learning becomes more popular in seizure analysis with large-scale datasets. In 2014, the American Epilepsy Society, Epilepsy Foundation of America, National Institutes of Health and Kaggle launched a seizure prediction competition together to predict seizure with a 1-hour lead time using seizure data from five canines and two humans [13]. The top 10 submitted solutions, using SVM, random forest, Convolutional Neural Networks (CNN), etc, achieved sensitivity at 75% and specificity from 0.33 to 0.75. [14] discovered a specific ictal pattern of channels with seizure activities in the time-frequency domain and called it a fingerprint of the epileptogenic zone. Their EZ-Fingerprint model predicted 64 contacts and 58 of them are inside of patients' resected areas. By using the resection zone as ground truth, their model achieved 90.6% positive predictive value and 0.7% false-positive rate.

According to the selection of training data, existing machine learning model on EEG signals can also be defined by three types: 1) Patient-specific model, which is trained only by the data from the testing subject; 2) inter-patient model, which is trained by the data from not only the testing subject but also other subjects' data; 3) cross-patient model, which is trained only by the data from other subjects. The cross-patient model is the most challenging one but also the most practical approach in the real world. A cross-patient model has potential of being immediately applied to new patients even if they do not have any collect EEG data in the system. The doctors may benefit from the automatic EEG signal classifiers since the models can help them to annotate the existing EEG recording or even supporting their decision making for diagnosis. Patients will also take advantage of seizure alert or seizure prediction applications. Besides, wearable devices for warning seizures have been developed and tested [15, 16]. Mobile devices compatible with seizure data offer a great opportunity for machine learning models to improve epilepsy patients' daily

lives. However, current machine learning projects on EEG signals are facing three challenges:

- Dataset limitations. Public datasets have limited cohort diversity because of the small number of subjects and from the same resources.
- Subjects variety. Seizure signals vary from patients but most algorithms are focusing on patient-specific and inter-patient models.
- Long-term Evaluation. Existing models lack reporting their results on continuous long-term EEG monitoring so the performance may differ from real-world scenarios.

1.4 Contribution

In this dissertation, we introduce an end-to-end pipeline combining data management and EEG signal analysis for epilepsy study. Figure 1.5 illustrates a conceptual diagram of this dissertation. We developed TeQ, a temporal events query system that extracts epilepsy temporal information from large-scale cross-site file system and provides a graphical query interface for EEG signal discovery. The following are three epilepsy research topics: EpiD for epilepsy detection, EpiP for epilepsy prediction and EpiL for epileptogenic zone localization. By using TeQ, we created datasets appropriately and performed robust evaluation for each study.

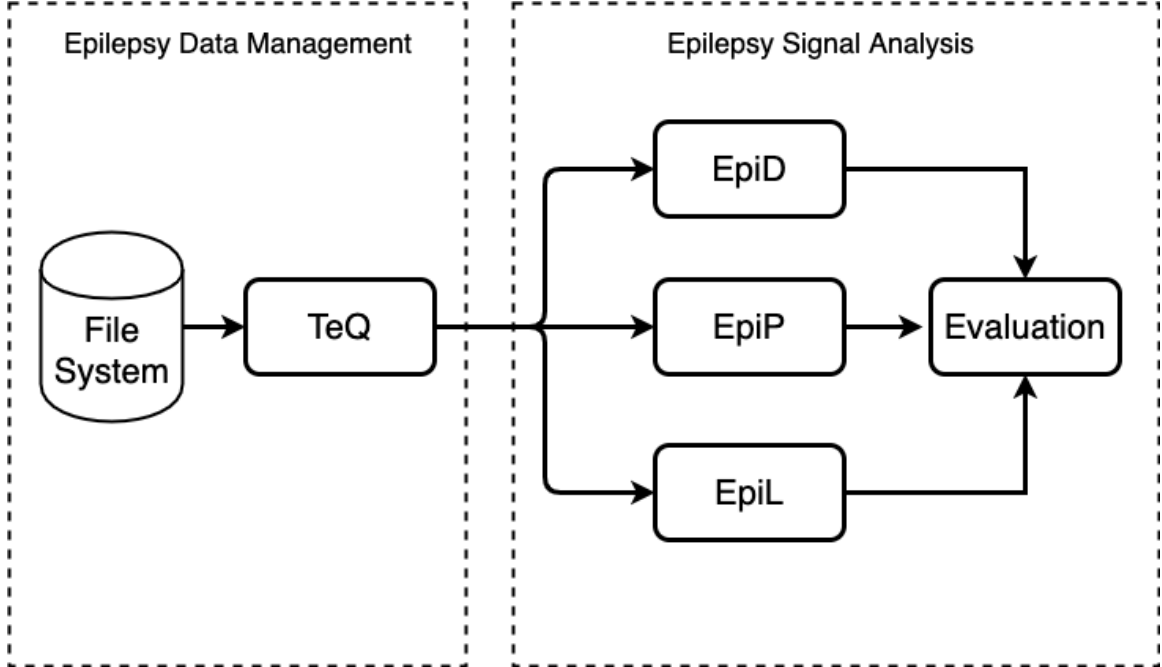


Figure 1.5: A concepts overview of this dissertation. TeQ: Temporal events Query. EpiD: Epilepsy Detection. EpiP: Epilepsy Prediction. EpiL: Epileptogenic zone Localization.

Comparing to existing methods, our work contributes to the following aspects:

Large-scale temporal information extraction. Our ontology-guided multi-site epilepsy temporal information system processed 2,497 epilepsy patients with 3169 reports from 7 epilepsy centers across the U.S. and Europe. We extracted 451,076 temporal annotations from 42,239 EEG files. We constructed vocabulary sets including 46 standard annotation terms for ontological annotation elements and matched 6,687 annotations for high-quality queries.

Prospective temporal data quality measurements. Our system prospectively integrates the epilepsy temporal data in CSR once a week. We automatically calculated the data quality measurements for the epilepsy temporal data. The results show the CSR dataset has 99.12% annotation completeness, 61.71% EEG signals completeness for all existing monitoring, and 0.85% signal file duplication rate.

Graphical temporal query. Our system provides a web-based temporal query interface developed by the RoR development framework. Both query widget and

results are displayed in the graphic. The temporal query canvas can generate all 13 Allen’s interval algebra with minimal user intervention. By using our interface, users can download the query results in the CSV format for preliminary research or datasets builds.

EEG signal classification performances. We developed three machine learning models for EEG signal classification. In the cross-patient case-based testing, our seizure detection model detected 90.75% lead seizures and achieved 92.23% overall accuracy, 93.57% specificity, 81.08% sensitivity, and 85.41% AUC in the segment-based evaluation. Our seizure prediction model reached 86.79% sensitivity and 3.38% false-positive rate. Our seizure localization model achieved 88.22% accuracy, 34.99% sensitivity, 1.02% false-positive rate, and 34.3% positive likelihood rate.

Long-term EEG evaluation. We built a long-term EEG evaluation dataset leveraging CSR large-scale EEG data volume. The sub-dataset has a high data quality with a 98.98% EEG signal completeness which is significantly greater than the 59.57% data completeness of CHB-MIT scalp EEG dataset. Our evaluation results are based on 2097 hours testing of our seizure detection model and 1506 hours of our seizure prediction model. The evaluation of our dataset is closer to a real-world situation, therefore, the results are more convincing.

1.5 Outline

This dissertation is organized as follows: In Chapter 2 we provide background knowledge on epilepsy data sources, multi-site epilepsy data management and EEG signal classification techniques and material for machine learning methods and development tools. Next, in Chapter 3, we introduce our ontology-guided multi-site epilepsy temporal data extraction and query system. In Chapter 4-6, we describe the details of developing our machine learning models for scalp EEG seizure detection, scalp EEG seizure prediction, and SEEG seizure localization. Finally, in Chapter 7, we

summarize the dissertation and provide direction for future work.

CHAPTER 2. Background

In this chapter, we describe the background knowledge and provide a literature review of the data, methods and technology related to the projects. We first describe four EEG signal datasets used for epilepsy research and EEG signal classification. Secondly, we review the principles and existing systems for cross-site epilepsy data management. Next, we provide background information of the machine learning technology used in this dissertation and existing work on EEG signal classification. Then we present the challenge of current evaluation data and methods and how we compare our models with existing work. At last, we acknowledge state-of-the-art development tools that we used in our implementation.

2.1 Epilepsy Data Sources

Since the first brain wave is recorded by Hans Berger in 1924 [17], both of its accuracy and convenience have been improved. A significant change is that the EEG data storage had evolved from paper and ink into standardized digital formats. The digitalization helps researchers using computers to share, visualize, compute and analyze a large amount of EEG data. Today, multiple EEG datasets are available to the public on the internet [18, 19]. Those datasets have been used in the research area of diseases (such as epilepsy, Parkinson’s, Alzheimer’s and depression), motor imagery, emotion recognition, etc.

2.1.1 Public Epilepsy Datasets

University Hospital Bonn Germany. There are three commonly used epilepsy dataset. First is the seizure dataset from University Hospital Bonn Germany which contains 500 23.6-sec single channel EEG fragments from 5 healthy volunteers and 5 patients [20]. The original recording used a 128-channel amplifier system and was

artificially selected to remove body movement noise. The written digital data was transformed from analog signals at a sampling rate of 173.61 Hz.

CHB-MIT. The second is Children’s Hospital Boston-Massachusetts Institute of Technology (CHB-MIT) Scalp EEG Database [21, 22]. The recording covered 182 seizures in 192 files from 22 subjects (5 males, ages 3-22; and 17 females, ages 1.5-19). The sample rate is 256Hz with 16-bit resolution and all the data contains at least 23 EEG channels (24 or 26 in a few cases) [23].

Kaggle. The third one is the dataset used in American Epilepsy Society Seizure Prediction Challenge on the data science competition platform kaggle.com (Kaggle, Inc. New York NY, USA) [13, 24]. EEG data from twelve subjects were provided in the contest. Eight of them were patients’ records that were collected by drug-resistant epilepsy undergoing intracranial EEG monitoring at Mayo Clinic Rochester; four of them were canines data from veterinary hospitals at the University of Minnesota and University of Pennsylvania. In total, 53 seizures were captured from patients and 42 were captured from dogs. Because of the benefits of intracranial EEG, the datasets were sampled in the rate of 400Hz(four dogs), 500Hz(one patient) and 5000Hz(seven patients).

2.1.2 Center for SUDEP Research

SUDEP is a life threatening disorder to people who frequently have seizures. Different estimates indicate that the risk of SUDEP varies from 0.2 to 2.7 cases per 1,000 person-year according to multiple analyzing methods and cohort [25]. Because the cause of SUDEP is unknown and the factors of seizure vary by patient, epilepsy remains a future danger to the patients. To understand the epilepsy related disease, such as SUDEP, it is urgent to build a large scale and rich informative epilepsy database with high data quality for researchers.

CSR is a National Institute for Neurological Disorders and Stroke (NINDS) funded

center without walls for collaborative research in Epilepsy. Composed of researchers from 14 institutions across the United States and Europe, CSR aims at the research area of SUDEP [26, 27] with extensive and diverse expertise. CSR provides a comprehensive, well-integrated retrospective repository of epilepsy-related data, consisting of bio-physiological signals linked to a risk factor and outcome data for participants in nearly 2,500 epilepsy patients. Being the largest and most comprehensive dataset in the epilepsy area, the CSR data has been manually curated with highest quality levels. It encompasses a wide variety of signals, data collection protocols, and processing algorithms, thus representing a significant but under-utilized resource of “big data.” CSR provides thousands of 24-hours rich annotated physiological signal recordings with European Data Format (EDF) files [28] of an enormous amount of epilepsy patients with a broad spectrum of age, social, racial, and ethnic.

Table 2.1: A comparison between CSR and other public datasets: University Hospital Bonn Germany seizure dataset, Children’s Hospital Boston-Massachusetts Institute of Technology, and 2014 Kaggle seizure prediction contest dataset. *Only includes the subjects with annotated seizures.

Dataset	UBSD	CHB-MIT	Kaggle	CSR
Subjects number	10	23	12	408*
Seizure number	100	182	95	1622
Sample rate (Hz)	173.61	256	400-5000	200-2000

2.2 Cross-site Epilepsy Data Management

The DIKW pyramid represents the structural relationship between data, information, knowledge, and wisdom. Each upper layer is the explanation or a higher level representation of the lower layer. In epilepsy study, each layer in the DIKW pyramid has

its own ability to answer different levels of questions from intuitive to difficult, they are:

- Data. Epilepsy data is the observation or measurement of objects or events. The data, such as signals, video recording, annotation text and symbols, displays the basic truth of the research targets.
- Information. Epilepsy information is the structural representation of epilepsy data, and it explains When, What, and Where information of the data.
- Knowledge. The knowledge of epilepsy domain can answer the “why” question. For example, the major question for a seizure is how it happens.
- Wisdom. With the existing knowledge, wisdom is developed to focus on the judgment of potential future events. For instance, forecast the probability of seizure in the next hours using current data.

An Information system (IS) integrates hardware and software to reduce the workload in the transformation from data to wisdom. IS is a network connecting the components such as data warehouse, database, user interface, and data management procedures. In particular, an epilepsy information system is a platform designed to integrate epilepsy data and extract epilepsy information, then it provides an interface for people to manage information, find new knowledge, and generate wisdom to prediction or prevent seizures. With the rapid increase of computational power, Artificial Intelligence (AI) added a new path for epilepsy exploration. Automatic seizure detection algorithms have been proven accurate with good performance but always come with a short latency. Practically, a seizure warning needs only a small period of time before the onset which provides huge protection to the patients. How to effectively and precisely predict pre-ictal periods before epileptic seizures occur using EEG data becomes a critical challenge. Thanks to the growth of big data and improvement of

AI technology, deep learning has become the most popular method for data analysis, classification, and prediction. Deep learning uses the idea of the human brain to build a machine neural network system that can learn and improve itself recursively from a dataset. In recent years, artificial neural networks achieved tremendous success in image recognition, natural language processing, recommendation systems, and many other areas.

While many machine learning models have been developed and widely used in seizure related topics, their performance on a real-world dataset is still questionable because lack of gold standard. In other words, the quantity and quality of the data may affect the results significantly for a machine learning model. At this point, a large scale and powerful epilepsy information system can play a important role. Its potential achievement includes:

- Ontology-driven EEG annotation integration: extract and map the annotation generated from different sources to a epilepsy ontology, and link them to EEG signals.
- Automatic epilepsy information import: insert or update information from newly added or curated epilepsy data, and provide scalability for data port in by new data vendors.
- Epilepsy data quality assurance: compute quantitative indicators of epilepsy data quality for datasets comparison and future improvement.
- Unified epilepsy information management: allow experts making changes on different sites using a single platform.
- Powerful temporal data retrieval: provide an effective human friendly temporal query interface to find a group of EEG segments with common event patterns.

- Customized epilepsy data preparation: store and output selected epilepsy data from certain cohorts with consumers’ demand.

In this section, we describe principles and existing systems for cross-site epilepsy data management.

2.2.1 Epilepsy Ontology

Ontology is the philosophical study of being, which describes the relations between concepts and basic categories of a domain knowledge. Today, ontology plays an important role in many information systems [29] and applications such as multi-site data integration, natural language processing, and decision support. In an epilepsy study, the annotations contain highly specialized epilepsy-specific terms or descriptions. An epilepsy ontology is a formalized terminology system to represent the knowledge of the epilepsy domain. With its help, relevant epilepsy information can be extracted and retrieved from these annotation data in free text. As part of the multi-center NINDS-funded study on sudden unexpected death in epilepsy (SUDEP), Epilepsy and Seizure Ontology (EpSO) has been developed for modeling highly specialized epilepsy and seizure-specific terms [30]. EpSO is used for multiple epilepsy domain applications, such as patient data entry, epilepsy focused clinical free text processing, and patient cohort identification.

2.2.2 FAIR Data Principle

With the rapid development of computational science, data sharing is becoming a primary feature in the big data community. According to FigShare’s latest annual open data report, 64% of survey respondents indicated their data was shared to the public during the year of 2018 [31]. During the whole life cycle of data science projects in average, mining data for patterns and algorithm enhancing only cost 13% of time, while 82% of time is spent on collecting datasets, cleaning/organizing data, and build-

ing training sets [32]. How to reduce both labor and time cost on data preparation is now a big challenge for a big data project. In March 2016, FAIR data principle was introduced by a consortium of scientists and organizations in the publication “FAIR Guiding Principles for scientific data management and stewardship” [33], which is a guideline to improve data sharing on findability, accessibility, interoperability, and reusability.

2.2.3 Cross-site Epilepsy Data Capture and Integration

Multi-Modality Epilepsy Data Capture and Integration System (MEDCIS) [2] is an ontology driven data entry and integration System for seven sites in the CSR program. MEDCIS aims to collect epilepsy data across multiple centers based on a shared ontology. In general, epilepsy data includes two types: phenotypic data and annotated long-term monitoring signals. Phenotypical data, which is the patient information captured by a web-based interface, contains patient demography, patient history, medication status, patient diagnosis, etc. Epilepsy signal data such as EEG and ECG are stored in EDF files, and epilepsy annotation data is extracted from free text content in clinical notes. MEDCIS also provides a web-based query interface to identify patient cohorts from a diverse source. For instance, a simple query is “Show all the female patients ages above 60 with generalized tonic clonic seizure from University Hospital at Case Western Reserve University and Northwestern Memorial Hospital at Northwestern University.”

SeizureBank [5] provides a cloud-based data repository which contains a large amount and high diversified seizure-related electrophysiological signal dataset, and an intuitive web-based system for managing, querying, exporting and visualizing seizure-related signal data. The features of SeizureBank lead to reduced time, space, and labor costs of seizure analysis on the large-scale dataset with an efficient data preparation pipeline and easy-to-use system for data management and visualization. The

majority of data scientists regard cleaning and organizing data as the least enjoyable work [32]; however, with SeizureBank, researchers no longer need to spend their time on data preparation and cleaning and could be more devoted to their major area of seizure analysis. SeizureBank is used in several seizure-related studies, including seizure subtype classification and seizure prediction research.

2.2.4 Data Quality Assurance

Data quality assurance is crucial for data reuse in clinical research. For a cross-site clinical data system, one issue of data quality is the variability in use of standardized vocabulary [34]. In CSR, the Ontology-driven Patient Information Capture (OPIC) system is developed for uniform electronic patient data capturing. OPIC provides an incorporated standardized terminology using EpSO for seven data sources. Ontology-guided Data Curation for Multisite Clinical Research Data Integration (ODaCCI) [35] introduces a streamlined data integration and curation workflow for CSR data quality assurance. Common data elements (CDEs), the data fields selected by epilepsy domain experts that are common to all individual clinical sites, are extracted by data source mappings. The system automatically computes completeness and consistency for all CDEs to evaluate the data quality of each site. Table 2.2 shows the data completeness report for ten CSEs in CSR. With the help of data quality assurance, an improvement of completeness has shown from 2016 to 2019 while the total patient reports number increased more than four times. However, the existing system only measured the quality of phenotypic data. To expand the capability, we introduce the data quality measurements for temporal epilepsy data in Chapter 3.

Table 2.2: Data completeness for ten common data elements in patient reports from multiple sites.

Concept	UH	NYU	UCLA	NW	TJU	UCL	IOWA	2019 Total	2016 Total
Age	99.51% 1618	98.58% 278	100.0% 125	100.0% 448	100.0% 225	98.63% 288	99.42% 170	99.46% 3152	97.76% 612
Gender	100.0% 1626	99.65% 281	100.0% 125	100.0% 448	100.0% 22525	100.0% 292	100.0% 171	99.97% 3168	92.52% 643
Drug	91.82% 1493	99.29% 280	97.6% 122	96.88% 434	97.33% 219	100.0% 292	94.15% 161	94.70% 3001	92.09% 640
Semiology	78.97% 1284	99.65% 281	52.80% 66	96.65% 433	96.89% 218	99.66% 291	72.51% 124	85.11% 2697	79.86% 555
Etiology	88.75% 1443	89.36% 252	8.80% 11	82.59% 370	10.67% 24	73.97% 216	9.36% 16	73.59% 2332	79.71% 554
EEG Type	93.05% 1513	13.12% 37	94.40% 118	81.92% 367	11.56% 26	86.3% 252	95.91% 164	78.16% 2477	75.97% 528
Epileptogenic Zone	72.94% 1186	95.04% 268	34.4% 43	54.46% 244	31.56% 71	99.32% 290	50.88% 87	69.08% 3152	70.65% 491
MRI/CT status	60.27% 980	73.76% 208	86.4% 108	91.74% 411	83.11% 187	97.6% 285	88.3% 151	73.52% 2330	69.35% 482
Ictal Seizure Type EEG	66.05% 1074	79.08% 223	73.6% 92	72.77% 326	0% 0	94.18% 275	63.74% 109	71.3% 2099	65.65% 411
Epileptiform Discharge	59.35% 965	86.88% 245	38.4% 48	66.74% 299	72.0% 162	86.64% 253	52.63% 90	65.07% 2062	57.99% 403
Total reports	1626	282	125	448	225	292	171	3169	695

2.2.5 EEG Information Retrieval

The last but not the least component of a cross-site clinical data system is information retrieval. How to search for specific signal segments from unstructured and high volume EEG data remains a challenge for cross-patient epilepsy study. Although Semi-structured temporal data, such as textuary EEG annotations, can help to describe the EEG signals, the information can only be retrieved when the related EEG signals file is read using existing tools. It is difficult to perform cohort discovery or preliminary analysis on cross-patient study using existing methods.

Current clinical data management system provides two types of temporal query. The first is structured temporal data query. For example, Research Electronic Data Capture (REDCap) provides a web-based interface for record query using a query form with defined input data fields. Users can set restriction on timestamps to search for records during a specific period. This approach is easy and fast but has functionality limitations on comprehensive temporal query, such as the temporal relation between two events. The second type of temporal query is syntax-based. “AMAS” is a temporal query language designed for the medical domain users to search and interpret clinical temporal data [36]. The syntax contains time and logic operator so the query is flexible to select patients who satisfy the temporal conditions. The learning process for new user is the challenge to be overcome for syntax-based temporal query. In Chapter 3, we introduce a new graphical temporal query interface, which is not only intuitive to use, but also powerful on temporal data query.

2.3 Machine Learning Methods

2.3.1 eXtreme Gradient Boosting (XGBoost)

XGBoost [37] is a scalable end-to-end machine learning system for gradient boosting machine [37]. The model of XGBoost is decision tree ensemble which consists of

a set of classification and regression trees. In practices, a single tree is not robust enough to solve a problem with high dimensional features, the solution is to combine the prediction of multiple trees. XGBoost provides a pipeline for fast and accurate training the tree ensemble model by parallel tree boosting. The system is a portable library for most supervised machine learning problems, even for a implementation on billions of examples training with memory-limited settings. In this dissertation, we used XGBoost to build a seizure detection model described in Chapter 4.

2.3.2 Convolutional Neural Network (CNN)

Inspired by biological neural systems in human brains, artificial neural networks are developed and widely used in data mining domain. CNN is a popular and powerful class of deep neural networks, it is commonly applied to the image classification problem. The ImageNet project is a large-scale hierarchical image database for use in visual object recognition machine learning research. More than 14 million images have been human-labeled by the project to annotate the objects in the pictures [38]. A trimmed 1000-category ImageNet dataset used in a famous annual competition is now known as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). VGG [39] is one of the winner model based on CNN. One improvement from previous models is using convolutional layers with smaller filter size (3X3), and the model can be improved by adding more layers. The disadvantage of VGG is the number of the parameters is large(>500 million) because the network is “very deep”, so training may take longer time than other models. ResNet [40], the winner of ILSVRC 2015, aims to learn the residual representation functions instead of learning the signal representation directly. Unlike the traditional stacked-layer models, it adds shortcuts in the network to make the learning more efficient. ResNeXt [40] is developed from ResNet by including the idea of inception: split, transform and merge.

2.3.3 Transfer Learning

Transfer learning is a machine learning problem that applies existing models on related tasks. A transfer learning method includes four steps: 1) select base dataset; 2) train the base model; 3) reload and modify the base model; 4) fine-tune the parameters by the target dataset. In the EEG signal analysis field, it has been used for EEG classification between subjects [41]. Studies also show that transfer learning can be applied to detect seizure with accurate and robust results [42, 43]. Transfer learning has two advantages: 1) reusing pre-trained model can significantly reduce the training time and the requirement of hardware because step 1 and 2 are skipped; 2) when target dataset is small, it can still achieve the good performance if the datasets are similar. In this project, we implemented a transfer learning experiments between image classification and EEG signals classification. The pre-trained dataset is the ImageNet challenge dataset, and the target dataset is the time-frequency data for scalp EEG and SEEG. The approach efficiently achieved high performance by leveraging state-of-the-art CNN architectures and the large pre-trained dataset. Our transfer learning models are described in Chapter 5 and Chapter 6.

2.4 EEG Signal Classification

About 3.47 million people in the U.S. are potentially affected by seizure everyday [1], the importance of seizure detection and prediction has been noticed by researchers. However, no significant sign can be captured by human eyes to make a prediction before the seizure occurs. In recent years, machine learning methods had been used in many areas and a lot of successful applications had been developed, such as speech recognition, natural language processing, and computer vision [44–46]. Machine learning provides a possible solution for a reliable prediction of seizure.

2.4.1 EEG Signal Data Processing

EEG signal data, a period of the recording of brain neurons activities, is now one of the most widely used methods in epileptic seizure studies. Many automatic techniques are developed to recognize different epileptic events by analyzing the EEG signals, for example, EEG-based seizure detection. EEG signal data processing is necessary before seizure detection algorithm development for two reasons:

First is the dataset unbalance for ictal and non-ictal. In the CHB-MIT database, one seizure occurred every five hours on average, and the number enlarged to every 15 hours for the data of randomly selected 23 patients in CSR. Most published works used downsampling to reduce the number of non-ictal samples. [47] [48] and [49] random selected a amount of non-ictal samples accordingly. [50] [51] [52] and [53] only selected non-ictal the closest to the seizures. [12] randomly extracted two non-ictal samples every one hour. Upsampling for ictal samples by the overlapped sliding window was also used in [53]. The training ictal and non-ictal ratio varies from 1:1 [47, 51, 52] to 1:12 [50].

The second reason is that EEG recording usually has a high sampling frequency. For the scalp EEG dataset used in this paper, the CHB-MIT dataset has 256Hz and the CSR dataset has 200 Hz. To fit the seizure detection model, feature extraction is implemented to large-scaled EEG signals for fast decision making. Two efficient methods are using the raw (time domain) signals [47, 49, 50] and frequency domain [47, 48, 50, 54]. Statistic measurements, such as mean, median, variance, skewness, kurtosis, etc., are used in such extraction methods. Some image-based features, a transformation of spectrogram to RGB data, for example, are also popular because of the rapid growth of deep learning. [12] used signal decomposition techniques including Empirical Mode Decomposition (EMD), Discrete Wavelet Transform (DWT) and Wavelet Packet Decomposition (WPD) to expand features in sub-band of signals. Autoencoder [51, 52] is another method to reduce the data size without

losing important signal features.

2.4.2 Seizure Detection

The goal of EEG-based seizure detection is to recognize a period of ictal activities with high accuracy and sensitivity and produce several false detections as low as possible. Thanks to existing EEG recording with labeled seizures, big data analysis using machine learning for automatic seizure detection had made big progress in recent years. An early study on EEG-based seizure detection only had 22% accuracy [6]. In 2018, perfect results (100% accuracy, 100% sensitivity and 100% specificity) were achieved by an inter-patient model [12].

Existing models' implementation varied by the machine learning techniques they used. [47] built a k-nearest neighbor (k-NN) model, a traditional machine learning cluster algorithm. [51] used another popular model: Support vector machine (SVM). In [12] four algorithms, k-NN, SVM, random forest (RF), and multilayer perceptron (MLP), were used to train with features. Their results claimed that SVM and RF are better than the other algorithm. MLP is a class of feedforward artificial neural networks, and [52] also used neural networks but on time-frequency features. Deep learning is the most state-of-the-art machine learning area, [49, 54] used a convolutional neural network (CNN). [48] combined CNN with another deep learning model long short term memory (LSTM), and [53] built a CNN model on features from different views (multi-view learning).

2.4.3 Seizure Prediction

Seizure prediction using EEG data started from 40 years ago. In recent years, machine learning methods had been used in many areas and a lot of successful applications had been developed, such as speech recognition, natural language processing, and computer vision [44–46]. Machine learning provides a possible solution for a reliable

prediction of seizure.

Another problem for seizure prediction is that researchers have very limited access to seizure datasets from real patients [55]. For example, in the 2014 kaggle seizure prediction competition, datasets provided are from 5 canines and 2 human subjects. Their goal is to build a classifier that can identify whether a clip of ten minutes EEG recording is a pre-ictal or inter-ictal segment. They defined pre-ictal as 65 minutes to 5 minutes period before seizure onset. In total, 505 teams joined the competition and submitted 17,856 classifications of the unlabeled test data. The top 10 submitted solutions, using SVM, random forest, Convolutional Neural Networks (CNN), etc, achieved sensitivity at 75% and specificity from 0.33 to 0.75.

Most deep learning seizure prediction methods train and evaluate models using the data of a single patient because a study shows that seizure prediction is a patient-specific problem [21]. However, in the real world, the seizure occurrence for a specific patient is rare and capturing enough seizure data, especially for a period before a seizure, is difficult. A practical solution is to use a sliding window to generate training samples. The disadvantage is sometimes the sliding step needs to be very small to balance the dataset. The consequence is the whole dataset is filled with repeated information, which may over-fit the model and reduce the prediction performance.

2.4.4 Epileptogenic Zone in Epilepsy

In 1993, Luders et al. [56] defined the epileptogenic zone as “the area of cortex that is necessary and sufficient for initiating seizures and whose removal (or disconnection) is necessary for complete abolition of seizures”. Epilepsy Patients will be completely seizure-free after removal of epileptogenic zone. However, the epileptogenic zone can not be certainly identified before the surgeries renders the patient seizure free, so locating the epileptogenic zone is still a major challenge in clinical practices. Another challenge is the epileptogenic zone has no direct preoperative measurement. Existing

clinical method requires multiple tests and presurgical evaluations to define epileptogenic zone as overlap of a list of cortical zones: irritative zone, ictal-onset zone, epileptogenic lesion, etc [57].

A machine learning pipeline to locate the epileptogenic zone using Stereoelectroencephalography (SEEG) signals was developed by Grinenko et al. [14]. The authors discovered a specific ictal pattern of channels with seizure activities in the time-frequency domain and called it a fingerprint of the epileptogenic zone. The pattern includes three characteristics: 1) sharp transients or spikes; 2) multi-band quick activity concurrent; 3) suppression of lower frequencies. To extract such features, they applied the Morlet wavelet transform to SEEG data near seizure onset. After filtering, ridge detection, and masking, they extracted or computed frequency, timing, and areas to describe the processed data. Finally, an SVM classifier was trained using a dataset consists of 17 patients' SEEG data. Their results showed the fingerprint patterns exist in 15 of 17 patients. Their EZ-Fingerprint model predicted 64 contacts and 58 of them are inside of patients resected areas. By using the resection zone as ground truth, their model achieved 90.6% positive predictive value and 0.7% false-positive rate. The limitations of current machine learning approach are: 1) The pipeline is not fully automatic because users need to manually mark features including start and end of the fast activities; 2) Some restriction on the shape of seizure. For example, the seizure must have gamma activities longer than three seconds. In Chapter 6, we introduce a fast and automatic deep learning approach for epileptogenic zone localization.

2.5 Evaluation of EEG Signal Classification

An EEG signal classification problem is to determine a period of EEG signals which belongs to a pre-defined class. Usually, the class number of a classifier for EEG signals is two. For instance, the classes for seizure detection are “is a seizure” and “not a

seizure”. After the classifier or model is trained, we need an evaluation to show its performance. In this section, we describe how to construct evaluation datasets and what are the common used evaluation measurements for EEG signal classification.

2.5.1 Evaluation Data

In a machine learning project, the collected dataset needs to be split into three parts for different purpose: a training set, a validation set and an evaluation (testing) set. The training set is used to fit the model and the validation set is used to tune model hyper-parameters while training. After training, the final model will be tested on the evaluation set. Both the validation set and evaluation set should be a complete unbiased set of training with no data leaking to each other. The difference is that the final model will benefit from the validation set so we recognize the validation set as part of the training set. The second column in Table 2.3 illustrates the split ratio between the train sets and evaluation set for 9 published seizure detection methods. Because of the limitation of data samples, recent works are using k-fold cross-validation, where k indicates the repeat number during the testing. If $k=10$, then the split ratio is 9:1 which means the model is validated 10 times and at each time the testing set is 10% of all data. When the evaluation is cross-patient on N subjects, it equals to a N-fold cross-validation with a split ratio of N-1:1.

Another feature of seizure related classification dataset is the imbalance of seizure and non-seizure. However, the machine learning implementation shows improved performance on balanced dataset [58]. As shown in the fourth column in Table 2.3, most seizure detection studies balance the distribution of seizure and non-seizure data to or close to 1:1. To achieve the ratio, they subsampled the non-seizure data by random selection or extracting a certain period of signals. As a result of subsampling, the evaluation set may have a side fact that the non-seizure data is only a small percentage in the EEG recording which limits the coverage of the evaluation. In

Table 2.3: The data collection method of 9 published seizure detection methods.

Method	Split	Non-SZ subsampling	SZ:Non-SZ	Non-SZ usage(%)
Kiranyaz et al. 2014 [50]	3:1	close to seizure	1:12	5.87
Fergus et al. 2015 [47]	4:1	random	1:1	1.99
Xun et al. 2015 [51]	1:1	close to seizure	1:1	2.19
Thodoro et al. 2016 [48]	N-1:1	random	1:4	0.40
Yuan et al. 2018 [52]	1:1	close to seizure	1:1	2.19
Zhou et al. 2018 [54]	5:1	x	x	x
Park et al. 2018 [49]	x	x	1.32:1	4.33
Alickovic et al. 2018 [12]	9:1	per hour random	1:2	3.1
Tian et al. 2019 [53]	4:1	close to seizure	x	x

this work, we also evaluate our model using continuous long-term EEG signals which provide significant larger coverage of the collected dataset.

2.5.2 Evaluation Measurements

After applying the final model to the evaluation set, we can build a confusion matrix to describe the complete performance of the model. For a binary classification problem, we have positive samples and negative samples. Four counts of sample numbers are listed in the matrix: 1) True positive (TP); 2) True negative (TN); 3: False positive (FP); 4) True negative (TN). We can calculate the following common used evaluation measurements based on the four counts:

- $Accuracy = \frac{TP + TN}{TP + TN + FP + FN};$
- $Precision = \frac{TP}{TP + FP};$
- $Recall (Sensitivity) = \frac{TP}{TP + FN};$
- $F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall};$

Considering the imbalance of positive and negative in the real-world scenarios, those measures may lead to false sense of achieving high accuracy, precision and

recall as listed in the example in Table 2.4. We need an evaluation on the proportion of negative samples that are mistakenly predicted as positive, so we have:

- *Specificity (False Positive Rate)* = $\frac{FP}{FP + TN}$;

In Table 2.4, the model’s specificity and sensitivity is always 80% and 100% whatever the positive negative ratio is. A receiver operating characteristic (ROC) curve is plotted by a group of sensitivity and specificity pairs at various thresholds. With the plotting of the ROC curve, we can calculate the Area Under Curve (AUC) as a combination of sensitivity and specificity.

Table 2.4: The evaluation measurements for the same model on different evaluation sets.

Evaluation Set	TP	FN	FP	TN	Accuracy (%)	Precision (%)	Recall (%)
Balance	5	0	1	4	90	83.88	100
Imbalance	5	0	9	36	82	35.71	100

During a continuous long-term monitoring, the False Positive Rate (FPR) or False alarm rate is also defined as the number of FP per hour in average. From a patient’s view, however, the numbers are not intuitive to the method’s performance when actually using it. In Figure 2.1, the seizure detection model A and B have three false alarms for each so their FPRs are the same, but their performances differ by the distance between the alarms. If the patient receives a one-hour special care every time the alarm raises, model A will cost about three times more unnecessary attention on the patient. To address this challenge, we introduce a case-based evaluation for seizure detection which in addition to traditional segment-based evaluation. We split the long-term EEG signals into multiple cases with duration no more than one hour. A non-ictal case does not contain any ictal data and an ictal case only contains ictal

data. A non-ictal case test passes only if no alarm occurs and an ictal case test passes if a alarm successfully triggered. Using the case-based evaluation, we can calculate the non-ictal pass rates from model A and B are 66.67% and 88.89%.

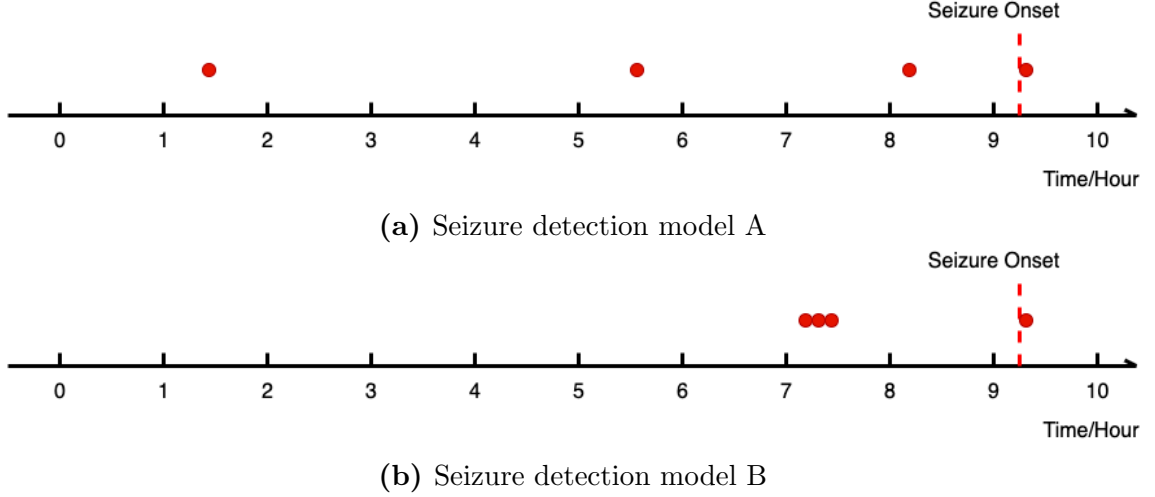


Figure 2.1: 10-hour continuous evaluations of seizure detection model A and seizure detection model B. Both of the models have three false alarms and one correct detection.

2.6 Development Environments

Our pipeline is built with modern computer software tools. In this section, we list three applications that are used for developing three major components: database for EEG metadata, data selection interface, and seizure prediction model.

2.6.1 Ruby on Rails

Ruby on Rails is a popular web development framework with two major guiding principles: “Don’t Repeat Yourself” and “Convention Over Configuration” [59]. “Don’t Repeat Yourself” is a principle in software engineering and is stated as “Every piece of knowledge must have a single, unambiguous, authoritative representation within a

system” [60]. By reusing code as much as possible, a Rails project is easy to extend the features, avoid bugs and reduce the work load of maintenance. The “Convention” of Ruby on Rails is a pre-build “Rails” for users to follow the best route of web development. The designed “Rails” make default settings rather than requiring manual configuration. Ruby on Rails is an open-source framework so its community is quite responsible and friendly. In addition, the programming language of Ruby on Rails is Ruby which is enjoyable to write and read. RubyGems, the package manager for Ruby programs and libraries, can easily extend the functionalities of application according to developers’ design. All the features of Ruby on Rails indicate its biggest goal is to help development to be faster and more efficiently.

Model-View-Controller (MVC) is an architectural pattern that is used for Ruby on Rails projects. Model is the kernel part of the architecture, it manages the logic and states of the application objects data. View is the upper layer of the pattern, it displays the information of model in particular representation and the interactive components to respond users’ behaviors. Controller is in the middle of Model and View to set up communication between them. Controller accepts input from users through View and applies commands for Model, then updates the representation of View accordingly. For example, my EEG annotation query tool has the structure of MVC. EDF file model stores the EEG metadata of each EDF files, and EEG annotation model stores the information of each annotation. The controller acquires user query variables including patient ID, annotation name, and duration from view and implements computation on model data and the View can display the query results on the webpage.

2.6.2 MySQL

MySQL is a free and open source relational database management system. It was original owned by the Swedish company MySQL AB and was acquired by Oracle

in 2010. MySQL is commonly used in data driven web applications because of its performance and flexibility. A database in MySQL is composed by a number of tables. In each table, a cell represents a data value, a row represents a single record, and a column represents a data field. Records can be identified by the primary key of the table, and tables are connected by foreign keys. MySQL also provides a query language which makes database operations including insertion, deletion, updating and searching more efficient. We use MySQL as our database of EEG metadata because: 1) Ruby programming language has the application programming interface (API) which includes a library of accessing MySQL, and Ruby on Rails uses MySQL as its default database system; 2) our data model does not contain complicated relation between tables so that a time consuming join query can be avoid; 3) As an embedded tool of MEDCIS, we use the same database system for a better query performance.

2.6.3 TensorFlow

TensorFlow [61] is one of the most commonly used machine learning platform. It is an open source software and was released by Google in 2015. As of Dec 2018, TensorFlow is the third most starred repository and has the fourth most folks on Github, the largest host of source code in the world [62]. TensorFlow includes a large amount of mathematical functions for artificial intelligence projects such as support vector machines and artificial neural networks. The basic component of TensorFlow is a directed acyclic computational graph which consists of tensors, nodes and edges. Tensor is a vector with N-dimension or a matrix of N rows in a simple way. The nodes represent operations and the operation results are tensors. Each edge represents the flow from an output of a node to an input of a node. In this project, we used TensorFlow to build deep learning and transfer learning and trained the model using my labeled EEG signal dataset.

CHAPTER 3. Temporal Query for Epilepsy Dataset

3.1 Motivation

Epilepsy study is broad and comprehensive involving many scientific disciplines and areas of exploration related to seizure disorder. Since the reason of seizure triggering are still unknown, collecting epilepsy data, such as Electroencephalography (EEG), magnetic resonance imaging (MRI), magnetic resonance spectroscopy (MRS), and positron emission tomography (PET), becomes necessary for future analysis. The number of epilepsy cases per year is more than 200,000 in the U.S. Although epilepsy is a common disease in the population (up to 10% of people worldwide have one seizure during their lifetime [63]), a seizure can vary in frequency, for example, it may happen once a year or several times per day. More importantly, a seizure typically lasts for less than two minutes, which causes the seizure data to be rarer compared to the non-seizure data. Some special cases related to epilepsy are much more difficult to capture. For example, SUDEP (about 1 in 1,000 people with epilepsy) patients are all fatalities so no more data can be captured after the patients being diagnosed with SUDEP. The solution is to collect data from multiple sites to increase the number of captured cases which significantly benefits researches on rare diseases including SUDEP. Nowadays, numerous universities and clinical centers worldwide are providing programs for epilepsy study. Leveraging epilepsy data collected from 7 epilepsy centers, scientists and physicians collaborated on The Center for SUDEP Research (CSR) to understand the rare and deadly disease. Informatics and Data Analytics Core for CSR is a majority component to build bridges between each institution and to provide an utilized ontology-based platform for data collection, curation, and sharing.

At each epilepsy center, An epilepsy monitoring unit (EMU) is an inpatient unit designed to evaluate, diagnose, and treat seizures by specialists. Depending on the

epilepsy patient’s status, the patient usually stays three to seven continuous days in the EMU. During a visit, the patient is performed long-term video-EEG monitoring, which continues for 24 hours or more. The patient can move, sleep, watch television, talk and participate in other normal activities while both of his/her EEG signals and video are recorded. The long-term video-EEG monitoring provides the connection between continuous behavioral observation and brain activities of epilepsy patients. The neurologists later can annotate the EEG data according to the signals and video. Because EEG signals and video files are storage-consuming and processing such large files is time-consuming, the annotation files for epilepsy data are an efficient way to describe a period of epilepsy recording especially for a large cohort or a multi-source dataset. Preliminary research can be done based on annotation information without acquiring and processing the complete large dataset. Furthermore, researchers may filter the dataset by selecting subjects or recording periods on specific study topics.

The existing epilepsy datasets, such as the CHB-MIT scalp EEG dataset, American Epilepsy Society Seizure Prediction Challenge dataset on Kaggle, and University of Bonn EEG dataset, are widely used for epileptic events classification. Both datasets from the American Epilepsy Society Seizure Prediction Challenge and the University of Bonn are highly prepared datasets for EEG analysis and do not contain dependent on annotation files. A well-prepared dataset is usually preprocessed by categories and split into small segments with constant length. University of Bonn EEG dataset contains EEG data from five non-epilepsy subjects and five epilepsy subjects with no temporal information but pre-labeled with five inter-patient categories: A) scalp EEG recording from the five healthy volunteers with eyes open; B) scalp EEG recording from the five healthy volunteers with eyes closed; C) intracranial EEG recording within the epileptogenic zone during interictal from the five epilepsy patients; D) intracranial EEG recording from the hippocampal formation during inter-ictal from the five epilepsy patients; E) intracranial EEG recording during ictal from the five

epilepsy patients. Users can retrieve the information by the data label from Table 3.1. The Kaggle dataset of the American Epilepsy Society Seizure Prediction Challenge used a similar way by describing the data using filenames and file directories. For example, filename “Patient_1_interictal_segment_0001.mat” indicates the EEG data of this file is recorded from 0 to 10 minutes from Patient1’s interictal period. The advantage of the two datasets is that they are ready to use and the information provided to researchers is clear and straight forward. Such datasets are built for specific purposes but limit the possibility of expansion.

Table 3.1: Retrieve information from University of Bonn EEG dataset using labels. The four characteristics are not in common for all labels and they are used to distinguish between each other.

Label	Subjects	EEG type	Location	Events
A	non-epilepsy	Scalp	Surface	eyes open
B	non-epilepsy	Scalp	Surface	eyes open
C	epilepsy	Intracranial	Epileptogenic zone	Interictal
D	epilepsy	Intracranial	Hippocampal formation	Interictal
E	epilepsy	Intracranial	With ictal activities	Ictal

CHB-MIT scalp EEG dataset is a lower level EEG dataset with minimum pre-processing so it obtained more original information from the monitoring. One type of information is temporal information which is important for time-series data, for example, EEG signals. In the CHB-MIT scalp EEG dataset, the temporal data includes file start time, file end time, seizure start time and seizure end time. Every subject has a summary file which stores the temporal data for each EEG file. From the four temporal data, we can compute multiple basic measurements for the EEG data, for instance, the duration of the total monitoring time, the duration of a seizure, the gap between files, the time distance of seizures, etc. By using the appropriate

measurements, we can extract specific data period from the whole dataset. One use case could be extracting all lead seizures by defining a lead seizure to be a seizure with no other seizures within one hour period before it. The temporal measurements can also be used for evaluating the quality of the dataset. For example, the long-term video-EEG monitoring should be continuous but a gap between two files is too long because of potentially missing files.

The size of three EEG datasets mentioned above is relatively small compared to the EEG data size in a real-world epilepsy center, which may contain hundreds of subjects. Besides, the dataset from an epilepsy center contains much more informative temporal data, especially annotation data. Since the temporal information of EMUs' EEG datasets is continuously growing, a temporal query system for large scale cross-site epilepsy datasets becomes useful for retrospective epilepsy research. However, there are four challenges:

- How to retrieve temporal information from multiple sources using different terminology;
- How to efficiently extract and store unstructured text annotation for information retrieval;
- How to build a user-friendly interface for temporal query purpose;
- How to intuitively display the query results for temporal data.

3.2 Dataset

The epilepsy dataset captured and integrated by MEDCIS from seven clinical sites, they are University Hospitals-Case Medical Center (UH-CMC), Ronald Reagan University of California Los Angeles (UCLA) Medical Center (RRUMC-Los Angeles), the National Hospital for Neurology and Neurosurgery (NHNN, London, UK), New

York University (NYU), Thomas Jefferson University (TJU), Northwestern Memorial Hospital (NMH Chicago), and The University of Iowa (IOWA). The dataset consists of two types of files: EDF files for signals recording (EEG, EKG, blood pressure, etc.), and the text files for annotations. All the data files are de-identified before storage and compressed into zip files. The details of file format and structure are described in the following:

Table 3.2: EDF header structure

EDF Header					
EDF File Metadata			EDF Signal Metadata		
Name	Type	Size (bytes)	Name	Type	Size (bytes)
Version	Integer	8	Label	String	16
Local Patient Identification	String	80	Transducer Type	String	80
Local Recording Identification	String	80	Physical Dimension	String	8
Start Date of Recording	dd.mm.yy	8	Physical Minimum	Float	8
Start Time of Recording	hh.mm.ss	8	Physical Maximum	Float	8
Number of Bytes in Header	Integer	8	Digital Minimum	Integer	8
Reserved	N/A	44	Digital Maximum	Integer	8
Number of Data Records	Integer	8	Prefiltering	String	80
Duration of a Data Record (seconds)	Float	8	Samples Per Data Record	Integer	8
Number of Signals	Integer	4	Reserved Area	N/A	32

Zip files – Zip is one of the most widely used file format for data compression. All epilepsy data files for each of the patient visit are compressed into a Zip file which named with a de-identified study ID (study identifier, or patient report identifier). A single zip file may contain multiple EDF file and annotation text file pairs.

EDF files – Epilepsy signal data is stored in EDF format. A EDF file consists of an EDF header and a series of signal records.

EDF header – EDF header represents the signal records metadata which can be split into two parts: metadata for the study, and metadata for each channel of signals. Table. 3.2 lists the EDf header elements details.

EDF signal data – Following with the EDF header is the EDF signal data that are stored in digits one after another in the order of signal channels listed in the EDF header. With start time of recording (t_s), duration of a data record (d), samples per data record (s), and the index of a signal in channel (i), the timestamp t of each signal point can be calculated by the formula $t = t_s + i \times \frac{d}{s}$.

Annotation text files – Each EDF file has a paired annotation text file which has the same base file name. In an annotation file, each row is a signal event with two elements: a relative time and a text annotation. The relative time can be used to calculate the absolute signal event time with the start time of recording from the EDF header. The experts manually reviewed and annotated the signal data with seizure events.

3.3 EEG Temporal Data Extraction

The general architecture of an ontology-driven information extraction system [29] is shown in Figure 3.1. We adapt the model to extract epilepsy annotation information from CSR data. The workflow is shown in Figure 3.2. The OPIC concepts are generated by EpSO for multi-site patient information capturing. The curation terms are provided by the experts who are annotating the epilepsy data. The combined terminology along with the specific task rules are used for matching the input annotation text. The input annotation text comes from annotation files with the time relative to the recording start time, which can be extracted from the EDF files. Each annotation time can be computed by the two-time values. Finally, the structural annotation records are generated and stored in the database.

By the reason that CSR is an on-going project and integrates new epilepsy data from its sites weekly, we build a rake task to automatically extract and store EDF and annotation information into the annotation database. Rake task enables developers to run their own Ruby code in a Rails project and to execute tasks periodically.

During the extraction process, the data quality dimensions can be measured for the input epilepsy data. The possible measurements include:

- completeness: determines if data is missing or unusable. For example, certain values in an EDF file are not available, or there is missing annotation file for an EEG recording;
- correctness: determines if data recorded the correct value. For example, the recording start time should be a specific time but not 00:00:00;
- consistency: determines if data conflicts with other data values. For example, a seizure phase appears after seizure end;
- duplication: determines if data repeats. For example, two same annotation records appear at the same time.

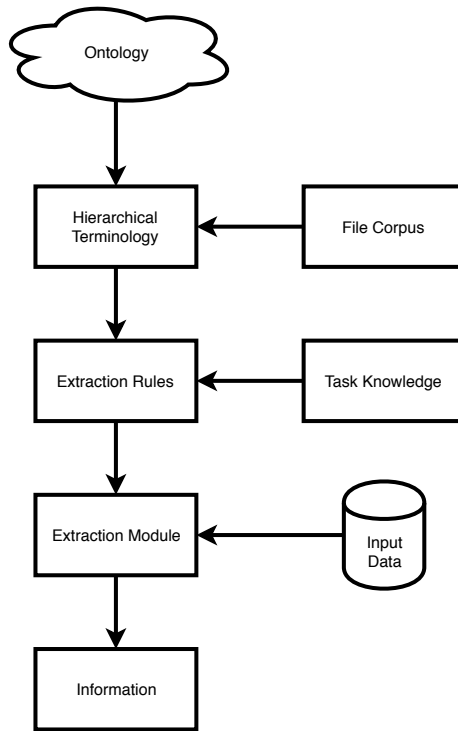


Figure 3.1: The general architecture of an ontology-driven information extraction system.

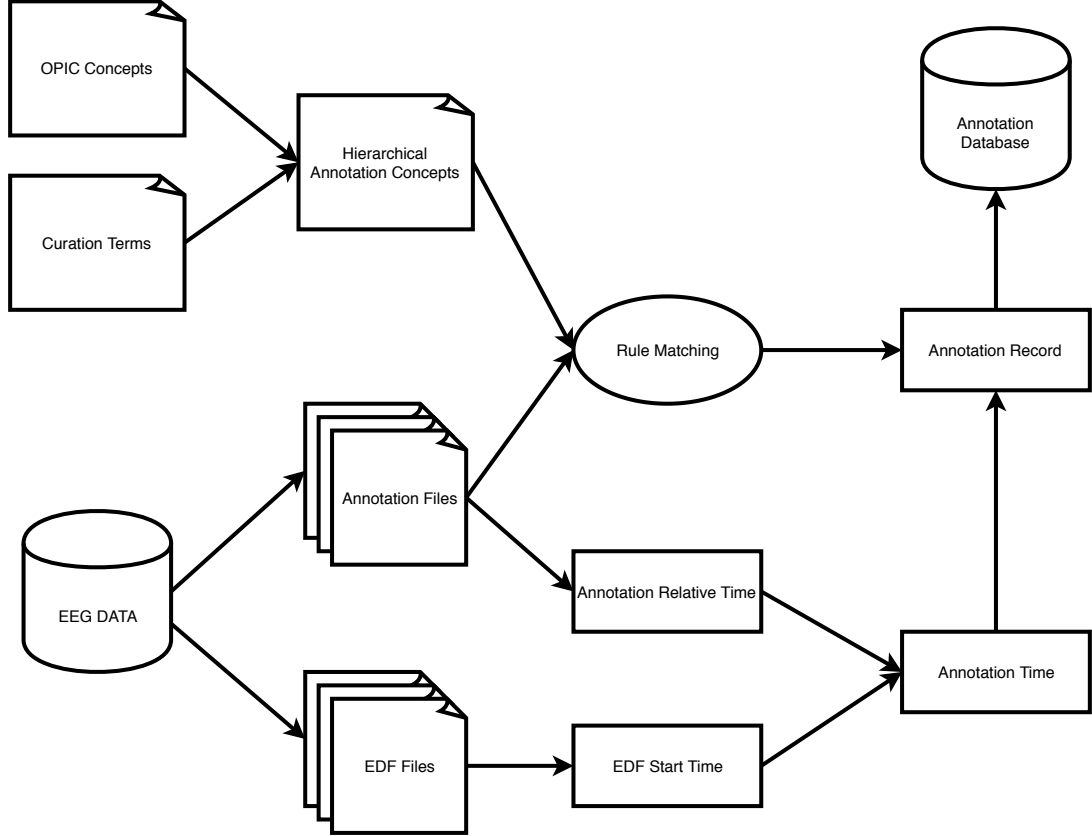


Figure 3.2: Ontology-driven annotation information extraction for CSR.

3.4 Temporal Query

In this section, we describe a full-stack method to query temporal annotation information. In general, temporal queries include timestamp queries and interval queries. A timestamp query retrieves all objects at a specific timestamp. An interval query retrieves all objects in a window of multiple consecutive timestamps. In most epilepsy researches, people want to find a specific annotation, for instance, find all “Clinical Seizure Onset” for patient A. The query is a single annotation query, which only implements a search in the annotation text dimension. Because our annotation system is ontology-guided, each concept has its unique ID. The feature transfers a single annotation query to an integer query instead of a string matching.

A more complicated situation is a multi-annotation query. For example, find

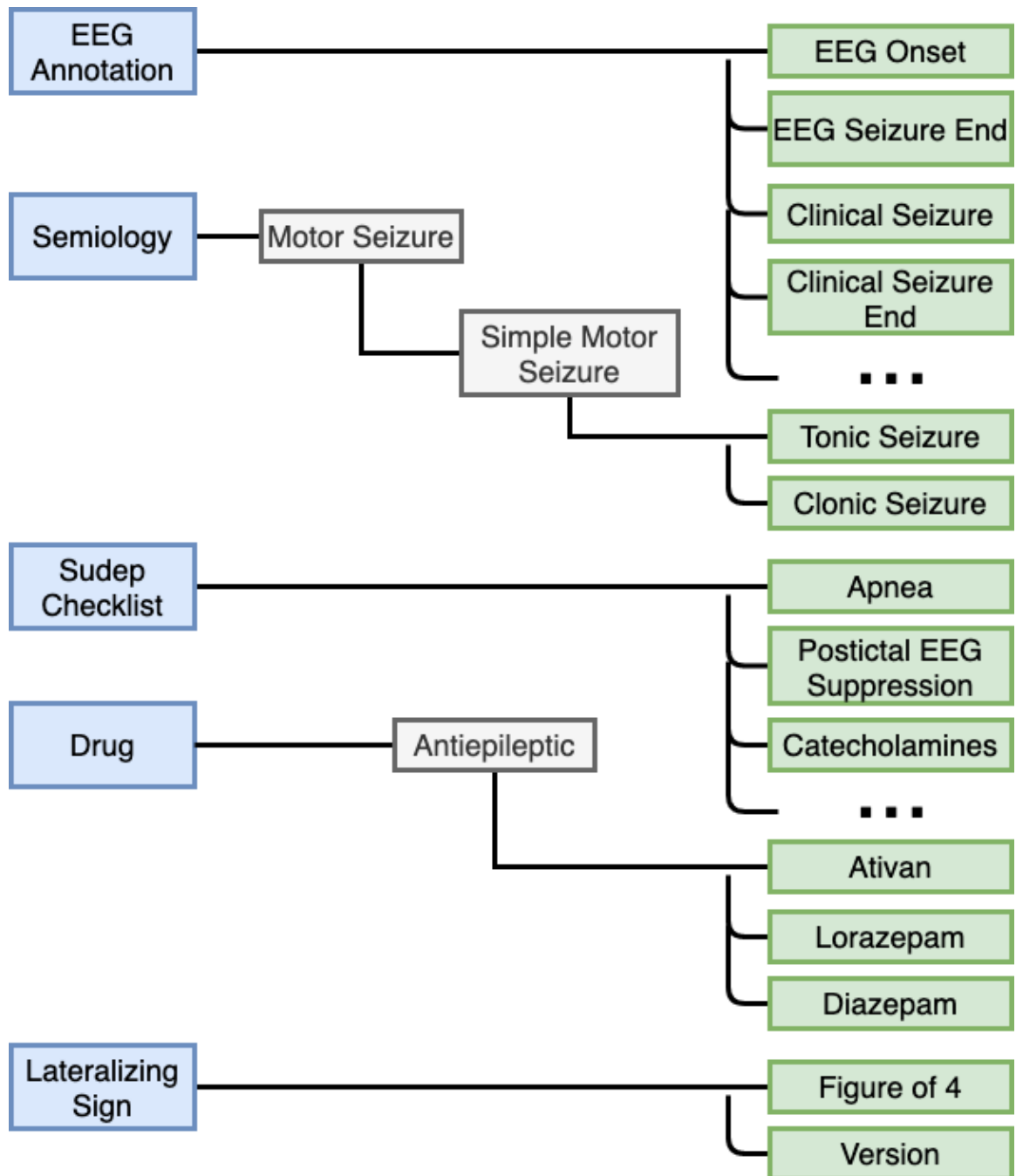


Figure 3.3: Ontology vocabulary for EEG annotation.

“Clinical Seizure Onset”, “Clinical Seizure End”, “Sign of Four” that “Sign of Four” is during “Clinical Seizure Onset” and “Clinical Seizure End”. In this case, not only concepts matching is needed, but also relations between concepts have to be considered. Allen’s interval algebra is a calculus for temporal reasoning that was

introduced by James F. Allen in 1983, the calculus includes 7 base relations: before, meet, overlap, starts, during, finish, equal. Since the time of an annotation can be a timestamp, we define a timestamp as a interval with the same start time and end time. Then, the system translate the query in human language to a syntax of Allen's interval algebra. In the multi-annotation query example, the syntax is:

{ Sign of Four } [during] ({ Clinical Seizure Onset } [before] { Clinical Seizure End })

The syntax is an infix expression which can be converted to an prefix expression:

{ Sign of Four } { Clinical Seizure Onset } { Clinical Seizure End } [before] [during]

Or it can be represented by a expression tree in Figure 3.4

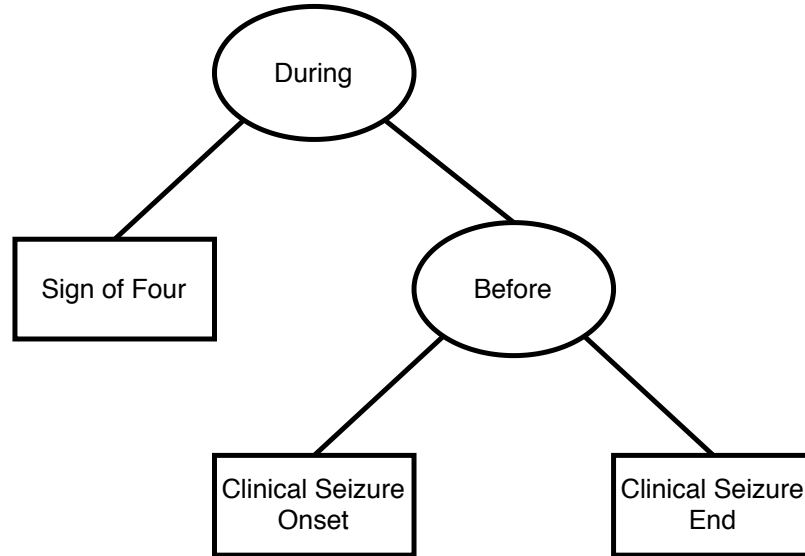


Figure 3.4: Expression tree representation for: “Sign of Four” is during “Clinical Seizure Onset” and “Clinical Seizure End”.

With the input of annotations and their time, a program can be implemented to tell if any combination satisfies the expression. The list of such annotation combinations is the results of the query. In epilepsy studies, the query could not only be a temporal query or an annotation query, but a two dimensional query. A two

dimensional query example is: find all annotations between “Clinical Seizure Onset” and “Clinical Seizure End” that “Sign of Four” is during such duration.

One of the two dimensional query topics is the spatio-temporal query, in which the two dimensions are space and time. Historical Rectangle Tree (HR-tree) is a data structure to resolve spatio-temporal problems. An HR-tree represents different versions of a R-tree at different timestamps. The advantage of the HR-tree is efficient timestamps query because the problem can be reduced to a R-tree problem. The disadvantage is that duplication of a whole rectangle is needed when one of the inside object changes, which increases the space cost.

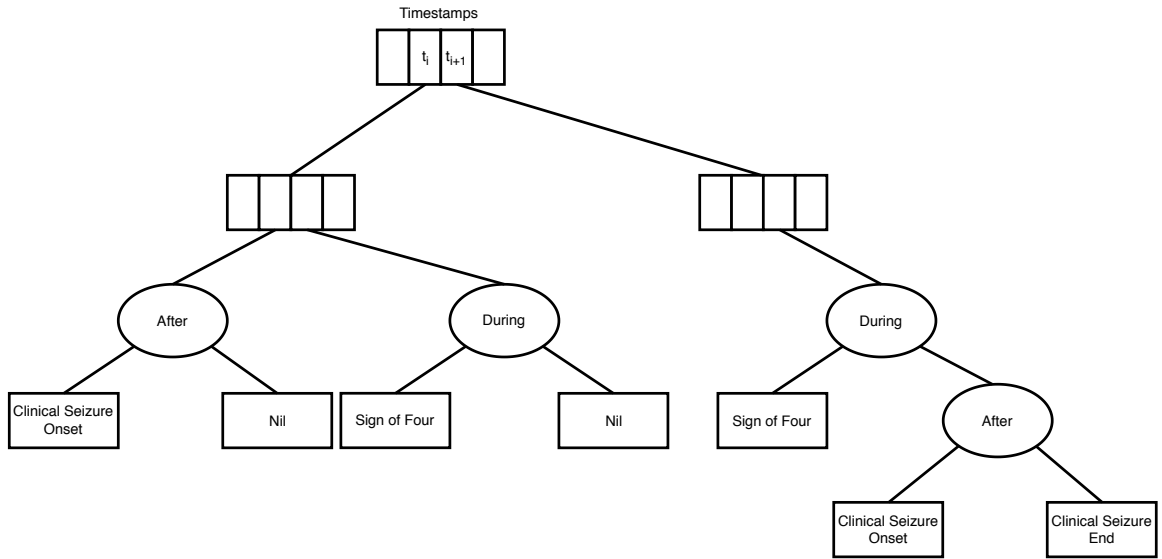


Figure 3.5: Example of a historical expression tree.

In order to efficiently use the epilepsy information, it is important to develop a human friendly user interface for the backend system. Our interface design could save researchers’ time on exploring the information system. Query is the major function for the interface. The query interfaces implemented in the following directions have three different input types:

- Syntax-based query, such as Google and database query language. The interface is simple because the only necessary input area is a text box. The query can

be powerful because users are able to modify most of the query, but building a query may be difficult if it is complicated.

- Structural query, such as shop websites. The inputs are usually given choices or limited to ranges according the value type. The developers can control the query inputs and avoid unexpected errors. The disadvantage is when the number of required input values is large, building a query may be inefficient.
- Graphic based, for example, graphic query interfaces for databases. Building query is intuitive for users because the query input exactly matches the results. However, more interface designing and development time may needed.

Our interface is a combined structural and graphical query interface and the graphic based query inputs are translated to Syntax-based query statements in the backend. We follow three principles for a information system user interface design:

- Efficient query builder. The number of clicks and typing for query input should be as small as possible;
- Informative results display. The results need to show all the information that users demand, and certain information can be found in a short time.
- Intuitive operation instruction. Without a user manual, the users can operate the functions by intuition.

As of January 2018, 1,309 study cases had been recorded and uploaded to the CSR system. We constructed a MySQL database for storing file information of 42,071 EDF files and 631,215 annotation records. My annotation query tool database design is shown in Figure 3.6. The EDF file information includes EDF file name, recording start date/time, and recording duration. It also contains two foreign key fields: patient study ID and database ID. Patient study ID links to the patient information in

MEDCIS, and database ID indicates the medical center where the recording comes from.

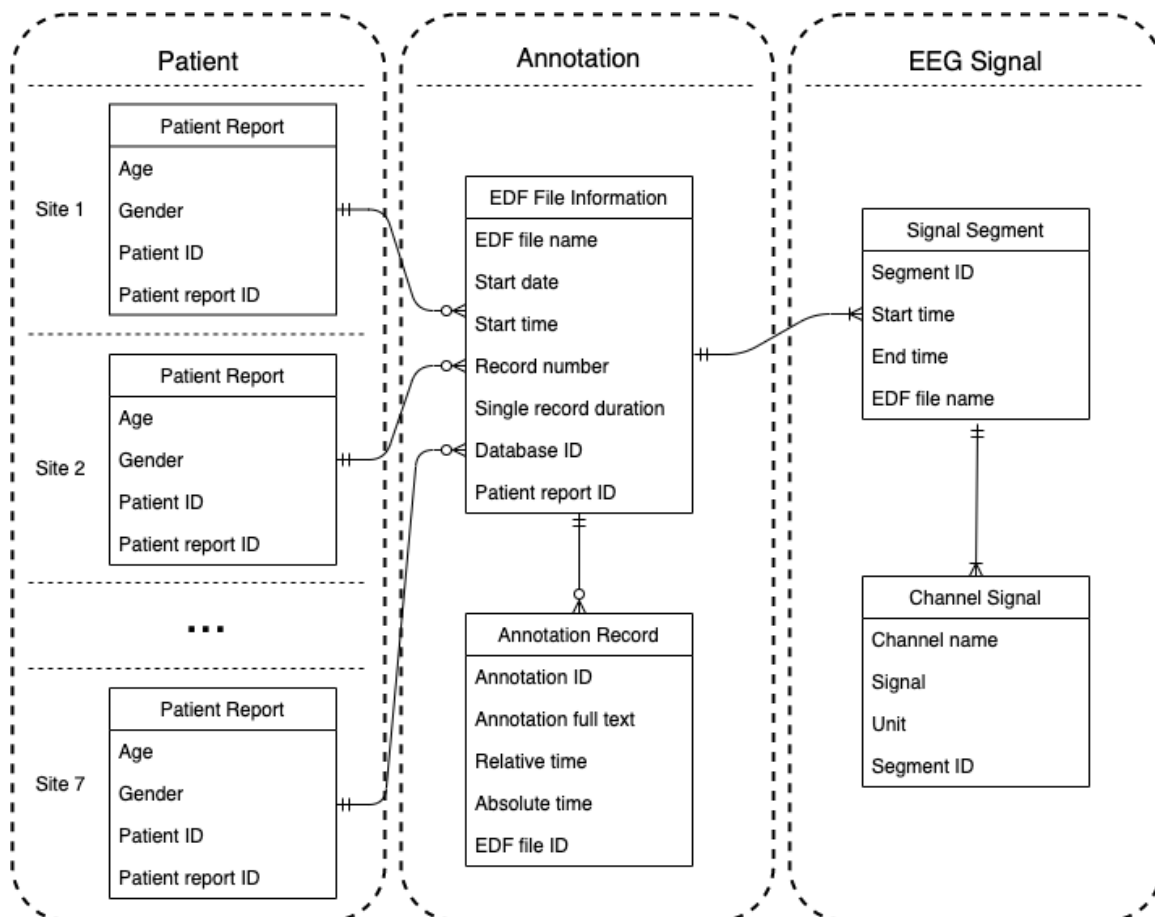


Figure 3.6: CSR temporal annotation query database design.

After the user finishes drawing the graphic query on the widget canvas, clicking the search button will start the query algorithm in the backend. The results will be displayed by subjects. The visualization of results will be very similar to the query widget, they will use the same rendering methods. Other features of the interface includes:

- Customized cohort: Users can select specific study subjects or groups for query;

- Annotation curation: Users can suggest improvement of the annotation data;
- Save and manege queries: Users can save a query for further review;
- Export query results: the results detail can be downloaded as a CSV file, and the users may use it for other analysis or processing;

3.5 Results

3.5.1 CSR Temporal Data Quality

In this study, we build a temporal query system on epilepsy data from seven clinical sites in CSR. The system connects to the CSR data integration platform and updates the temporal data weekly. We report the results using the September 4th, 2019 version that recruited 2497 patients and 3169 reports. The report number is larger than the patient number because a patient may visit an epilepsy center for more than one time. Among the 2497 patients, 1076 are from UH, 271 from NYU, 125 from UCLA, 393 from NW, 225 from TJU, 189 from UCL, and 170 from Iowa. Table 3.6 lists the EDF files completeness of patients and reports by different sites. The measurement indicates how many patients and reports have related EEG signal files uploaded to the system. As of September 4th, 2019, 125 patients from UCLA have no signals data so we did not include UCLA in the results of EDF file-related data quality measurements. TJU has the highest completeness rate of 74.07% by patients and 63.56% by reports and UH has the largest EEG signal dataset which contains 754 patients.

Table 3.4 indicates the data quality for collected EEG signal files in CSR. Almost all existing EEG signal files have attached annotation files with 99.12% completeness. However only NW and IOWA achieved perfect or nearly perfect completeness on recording start time for EDF files. Other sites have about 20% files with missing datetime information. Besides completeness, only six EDF files of all have incorrect

Table 3.3: EEG signal files completeness for patients and patient reports from multiple sites

Measurement	UH	NYU	UCLA	NW	UCL	TJU	IOWA	Total
By Patients	70.07%	48.71%	0.0%	39.19%	74.07%	24.0%	44.12%	52.42%
	754/1076	132/271	0/125	154/393	140/189	54/225	75/170	1309/2497
By Reports	54.43%	47.16%	0.0%	35.49%	63.56%	21.98%	43.86%	45.91%
	885/1626	133/282	0/125	159/448	143/292	60/273	75/171	1455/3169

start datetime in their header.

Table 3.4: Data quality measurement for EEG signal files from multiple sites

Measurement	UH	NYU	NW	UCL	TJU	IOWA	Total
Annotation	98.29%	100.0%	99.99%	100.0%	100.0%	100.0%	99.12%
Completeness	21505	1564	9659	1551	7265	695	42239
Datetime	73.74%	83.89%	99.99%	76.21%	78.79%	100.0%	81.44%
Completeness	16134	1312	9659	1182	5724	695	34706
Datetime	73.72%	83.89%	99.99%	76.14%	78.79%	100.0%	81.43%
Correctness	16129	1312	9659	1181	5724	695	34700
Total Files	21880	1564	9660	1551	7265	695	42615

We extracted 451,076 temporal annotations from 42239 files. 6687 of them matched 46 standard vocabulary in our ontological annotation list that are used for data curation. As shown in Table 3.5, four sites (NW, TJU, UCL, IOWA) have zero error rate on these curation annotations. The major two errors types are typos and merged annotations. The annotation duplication is displayed by a proportion of repeated annotation work which may have done on about 6% annotations.

In table 3.6, we computed completeness and duplication on every patient’s mon-

Table 3.5: Data quality measurement for EEG annotations from multiple sites

Measurement	UH	NYU	NW	UCL	TJU	IOWA	Total
Annotation	92.86%	94.33%	100.0%	100.0%	100.0%	100.0%	95.04%
Correctness	4172	183	691	29	386	894	6355
Standard Annotations	4493	194	691	29	386	894	6687
Annotation	96.75%	97.95%	82.97%	99.98%	95.99%	99.42%	94.53%
Duplication	228555	38966	66400	16876	58107	17518	426422
Total Annotation	236231	39781	80030	16880	60534	17620	451076

itoring and combined the results by each site. UH and NW have $\geq 80\%$ valid EEG signals coverage during the monitoring. UCL and IOWA have lower than 5% coverage. All sites have low duplication rate on stored EEG signal files, the highest is NW with 5.92%.

Table 3.6: Long-term EEG monitoring data completeness and duplication

Measurement	UH	NYU	NW	TJU	UCL	IOWA	Total
Completeness	80.1%	67.76%	84.69%	51.25%	4.43%	4.79%	61.71%
	2222.9	203.26	446.19	97.95	42.51	6.65	3019.47
Duplication	0.29%	0.17%	5.92%	0.99%	0.0%	0.0%	0.85%
	7.93	0.5	31.2	1.89	0.0	0.0	41.51
Total Monitoring Days	2775.1	299.97	526.87	191.13	960.78	138.86	4892.71

3.5.2 Temporal Query User Interface Design

In the frontend, we build a web-based query interface using the RoR development framework. The screenshot of the graphic temporal annotation query webpage is displayed in Figure 3.7. The interface consists of six components: 1) label of current

patient number in the system; 2) site selection; 3) temporal query widget; 4) results summary; 5) button for downloading the results in a CSV file; 6) timeline display of results by patients.

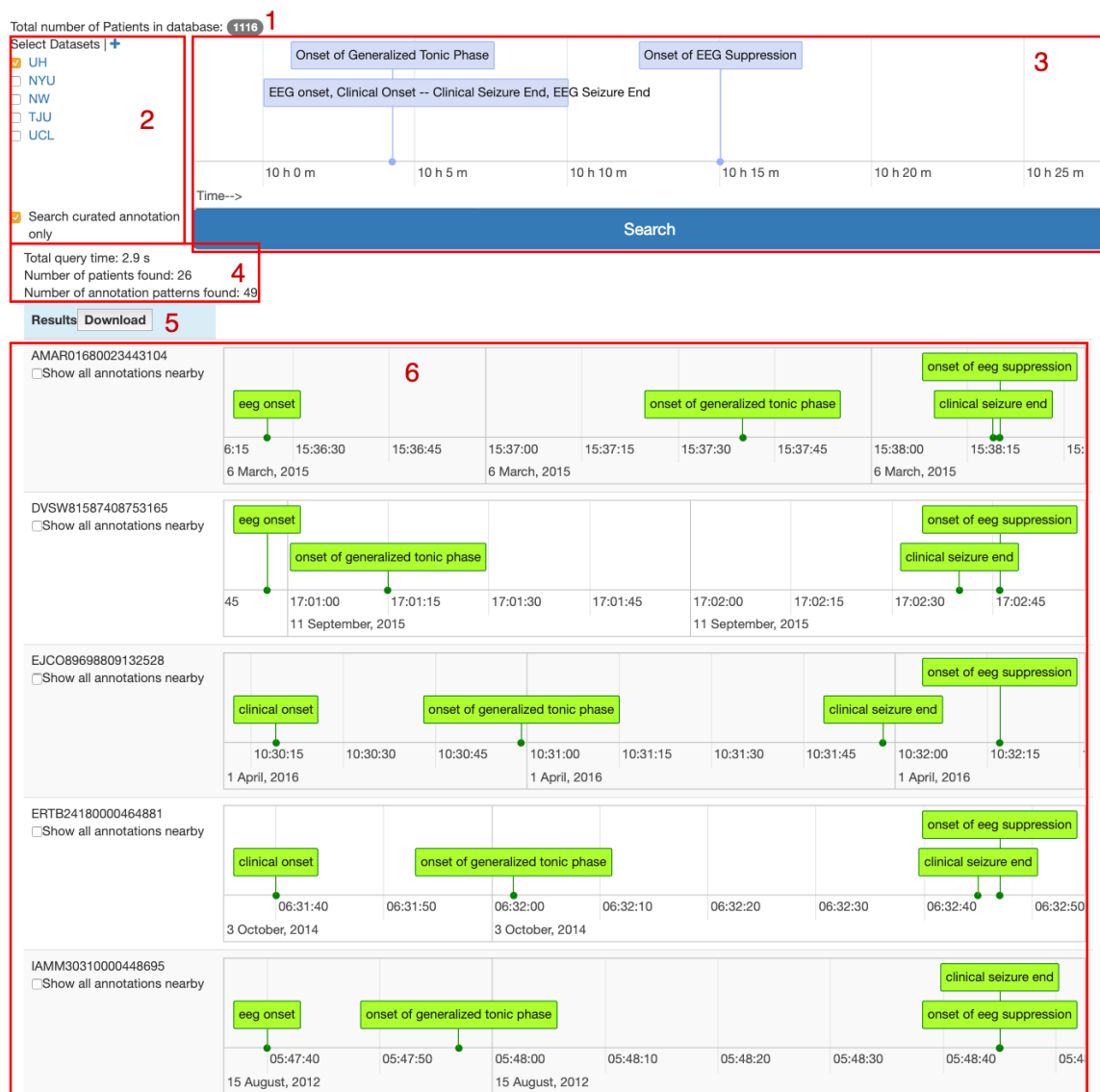


Figure 3.7: CSR Temporal annotation query interface.

The core of the interface is the temporal query widget. The widget displays a timeline where the time increases from left to right. Two types of temporal annotation can be added to the widget: timestamp annotation and interval annotation. The timestamp annotation is a box that represents a single name: “Onset of Generalized

Tonic Phase” and “Onset of EEG Suppression” in the example. The pointer marks its position in the timeline. The interval annotation is a box with a start annotation on the left and an end annotation on the right. In this example, the interval means a pattern starts from “EEG Onset” or “Clinical Onset” and ends with “Clinical Seizure End” or “EEG Seizure End”. The left edge of the interval represents the start time and the right edge indicates the end time of the interval. The position of two boxes expresses the relation of the annotations, and users can change the position to build all seven types of Allen’s interval algebra. In the query widget, the functions for users include:

- Add: double click the empty area, then a pop-up window will let users input the information of the annotation;
- Select: click an annotation box will select the annotation;
- Delete: click the red cross after selecting an annotation will remove it from the timeline;
- Edit: double click an annotation will let users edit the annotation information in a pop-up window;
- Drag: move when mouse is down on an annotation will change the position of the annotation;
- Zoom in/out: users can change the scale of the timeline window.

The users can select one or more pre-loaded subset from seven sites in CSR, or they can build their subset with patient ID by clicking the blue plus button. When the user finishes creating the query pattern and clicks the “Search” button, area four will display the summary of the results include total query time, number of patients have the query pattern and the total number of the match pattern. Division six shows

the timeline of detailed results for each patient with a pattern. Three functions are implemented for the resulting timeline:

(1) **Zoom out.** Users can enlarge the scale of the timeline. Figure 3.8 shows an example of the zoom out function. The patient has four matched patterns from July 28th, 2013 to August 1st, 2013.

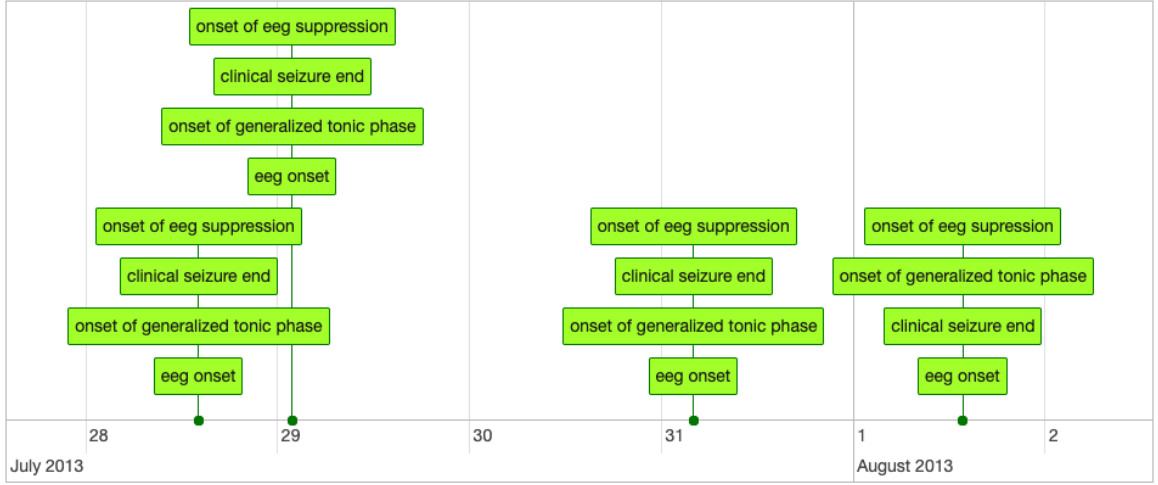


Figure 3.8: An example of zoom out function.

(2) **Zoom-in.** Users can shorten the scale of the timeline to see the detail of a pattern. Figure 3.8 shows an example of the zoom-in function. The window size is five seconds and the pattern lasts for about 2 seconds.

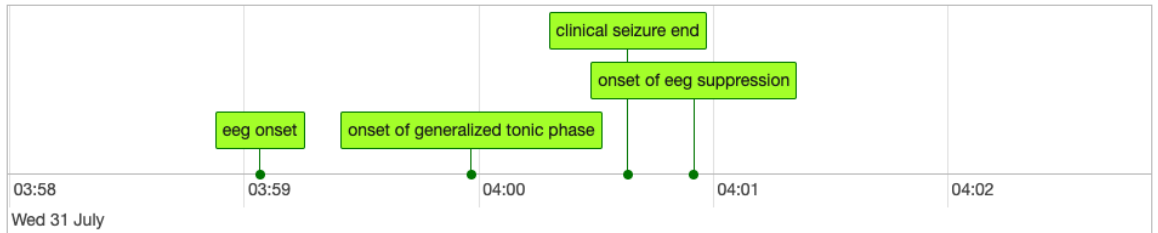


Figure 3.9: An example of zoom in function.

(3) **Show all annotations nearby.** Users can enable the display of all annotations in the period of the current time window. As shown in 3.10, if the user only queries

standard annotations, it will only display the nearby annotations that match the vocabularies in the ontology. As shown in Figure 3.11, if the user only queries all annotations, it will display all the nearby annotations in the database.

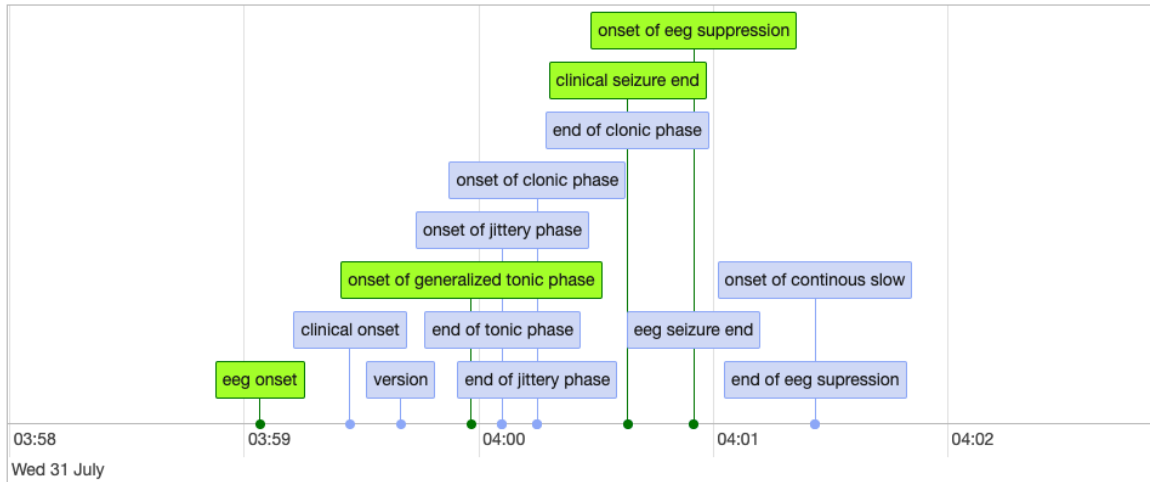


Figure 3.10: An example of showing all standard annotations.

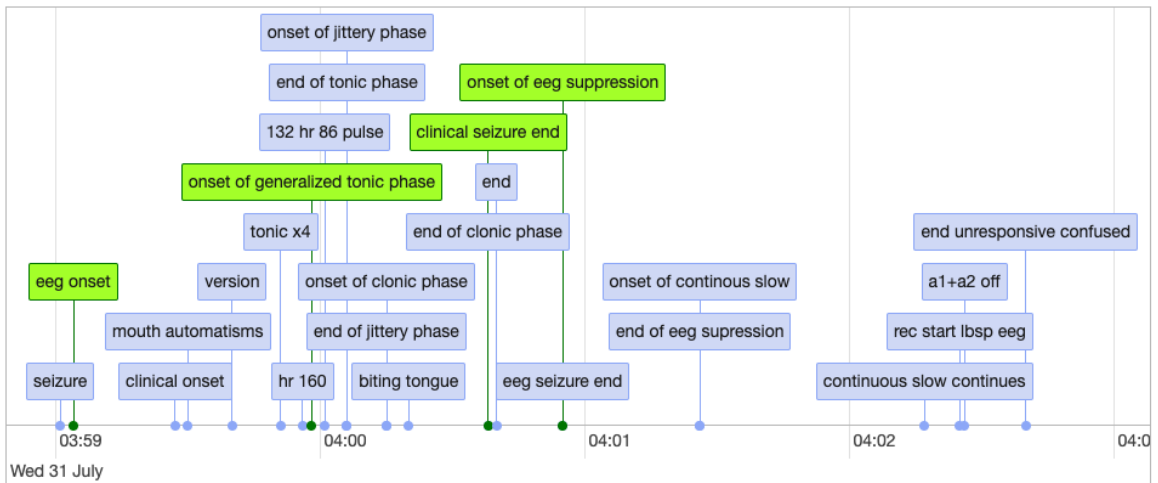


Figure 3.11: An example of showing all annotations.

3.5.3 Customized Epilepsy Dataset Builder

Users can download the query results in the CSV format using the button at the red number five in Figure 3.7. The file includes selected annotations and patterns with their DateTime, the stored file, and the relative time in the file. Another function is adding a period before or after the pattern and the interface can automatically indicate the part of file which contains the data by demands. An use case example of the temporal query interface is the seizure detection study we will introduce in the next chapter. We want to find all patients with annotated start and end of seizures, and we need the data to be as complete as possible so we can simulate a long-term EEG monitoring testing. As a result, we randomly selected 23 patients that match our requirements as our dataset. Table 6.1 shows the data quality of completeness of our dataset and a widely used dataset CHB-MIT scalp EEG dataset. Our dataset contains about 500 more recording hours and has 98.98% completeness, which is significantly greater than the data completeness of the CHB-MIT scalp EEG dataset. The evaluation of our dataset is closer to a real-world situation so the results are more convincing.

Table 3.7: The completeness comparison between the CHB-MIT scalp EEG dataset and a 23-patient subset of the CSR dataset.

Subject	Completeness	Data Duration	Monitoring Duration	Subject	Completeness	Data Duration	Monitoring Duration
CHB-1	89.03%	40.55 h	45.55 h	CSR-1	98.93%	112.46 h	113.68 h
CHB-2	86.92%	35.27 h	40.57 h	CSR-2	99.95%	66.12 h	66.15 h
CHB-3	46.42%	38.0 h	81.87 h	CSR-3	99.95%	94.85 h	94.9 h
CHB-4	96.17%	156.06 h	162.27 h	CSR-4	99.47%	65.69 h	66.04 h
CHB-5	99.79%	39.0 h	39.09 h	CSR-5	99.97%	139.63 h	139.67 h
CHB-6	74.77%	66.73 h	89.25 h	CSR-6	95.34%	110.36 h	115.75 h
CHB-7	99.72%	67.05 h	67.24 h	CSR-7	99.94%	120.47 h	120.54 h
CHB-8	75.84%	20.01 h	26.38 h	CSR-8	99.93%	49.04 h	49.07 h
CHB-9	99.56%	67.87 h	68.17 h	CSR-9	99.43%	71.02 h	71.43 h
CHB-10	29.69%	50.02 h	168.49 h	CSR-10	99.41%	74.32 h	74.76 h
CHB-11	35.76%	34.79 h	97.3 h	CSR-11	98.53%	77.46 h	78.62 h
CHB-12	70.84%	23.69 h	33.45 h	CSR-12	98.93%	70.47 h	71.23 h
CHB-13	54.43%	33.0 h	60.63 h	CSR-13	96.67%	166.91 h	172.66 h
CHB-14	61.79%	26.0 h	42.08 h	CSR-14	99.95%	44.49 h	44.51 h
CHB-15	63.21%	40.01 h	63.3 h	CSR-15	100.0%	44.6 h	44.6 h
CHB-16	99.82%	19.0 h	19.03 h	CSR-16	98.11%	75.3 h	76.75 h
CHB-17	24.09%	21.01 h	87.2 h	CSR-17	99.97%	132.0 h	132.04 h
CHB-18	40.55%	35.63 h	87.88 h	CSR-18	99.95%	120.39 h	120.46 h
CHB-19	33.06%	29.93 h	90.53 h	CSR-19	99.89%	118.06 h	118.19 h
CHB-20	42.18%	27.6 h	65.44 h	CSR-20	99.95%	99.8 h	99.86 h
CHB-21	58.93%	32.83 h	55.71 h	CSR-21	99.55%	70.53 h	70.85 h
CHB-22	40.83%	31.0 h	75.93 h	CSR-22	96.83%	117.36 h	121.2 h
CHB-23	56.62%	26.56 h	46.9 h	CSR-23	99.87%	56.24 h	56.31 h
CHB-Total	59.57%	961.64 h	1614.24 h	CSR-Total	98.98%	2097.57 h	2119.27 h

3.6 Discussion

In this chapter, we present a temporal data extraction and query system for cross-site epilepsy dataset. Because the input data of the system has a commonly used format, like EDF for EEG signals and TXT for annotation files, it is adaptable to other EEG dataset for temporal information extraction, data quality measurements, and epilepsy events query. The standard query term dictionary is scalable for specific research topics by adding new terms to a terminology file, the system can automatically update the ontology structure. Our system is the first among epilepsy data systems to provide a graphical temporal query interface, which is a fast and accurate solution for cohort discovery and pattern discovery on large scaled CSR epilepsy data.

Since epilepsy data collection in CSR is an on-going project, the completeness shows differently for the seven individual sites in Table 3.6. We also found mismatching problems for EDF files and annotations files. If the naming formats of EDF and annotation files are inconsistent, the pairing will fail which results in missing signal files or missing annotation files. The curation based on the data quality measurements will correct the errors in the future.

Another limitation of our work is that we extracted 46 standard annotation terms for curation and matched total 6,687 annotations from the dataset. It is only a small proportion of total 451,076 pieces of annotations we collected. In the interface shown in Figure 3.7, we add a button in area two that can disable “search curated annotation only”, so the users can still query all the free-text originally stored in the annotation files. Search by string matching will cause longer query time and the quality of the free-text is unknown.

3.7 Conclusion

Our ontology guided multi-site epilepsy temporal information system processed 2,497 epilepsy patients with 3,169 reports from seven epilepsy centers across the U.S. and Europe. We extracted 451,076 temporal annotations from 42,239 EEG files. We constructed vocabulary sets including 46 standard annotation terms for ontological annotation elements and matched 6,687 annotations for high-quality queries. Our system prospectively integrates the epilepsy temporal data in CSR once a week. We automatically calculated the data quality measurements for the epilepsy temporal data. The results show the CSR dataset has 99.12% annotation completeness, 61.71% EEG signals completeness for all existing monitoring, and 0.85% signal file duplication rate. Our system provides a web-based temporal query interface developed by the RoR development framework. Both query widget and results representation are displayed in the graphic. The temporal query canvas can generate all 13 Allen's interval algebra with minimal user intervention. By using our interface, users can download the query results in the CSV format for preliminary research or building their datasets.

CHAPTER 4. Seizure Detection on Scalp EEG Data

4.1 Motivation

According to the 2013 Institute of Medicine (IOM) report, 27% patients with poorly controlled epilepsy have accidents and injuries in a two-year period. Fatal accidents and injuries are the reasons for 6 and 20 percent of all deaths of people with epilepsy [64]. A system for seizure alert or seizure detection may prevent such accidents and injuries by notifying others when a seizure happens. A potential application of a reliable seizure detection tool will be long-term monitoring of seizures, whether using wearable devices or in the Epilepsy Monitoring Unit (EMU). By the reason that the non-seizure period takes the majority of time during long-term monitoring, a reliable seizure detection not only means high seizure detection rate (sensitivity) but also requires a low false alarm rate to prevent patients from panic and keep them relax in their daily routines.

For decades, brain activities of epilepsy patients have been systematically captured by monitoring patients' electrophysiological signals including Electroencephalography (EEG). EEG records voltage fluctuations resulting from ionic current within the neurons of the brain [3]. The long-term EEG monitoring is the EEG data recorded continuously for a long period (the average recording time per subject for the two used datasets are 41.8 hours and 91 hours). Using such long-term EEG signals, especially in the evaluation, removes the barrier between technical development and clinical implication in the EEG research area. A long-term EEG monitoring can be split to 4 phases by seizures: 1) The ictal phase is the duration between the start of a seizure and end of the seizure; 2) The pre-ictal phase is the period before the start of a seizure; 3) The post-ictal phase is the period after the end of a seizure; 4) The inter-ictal phase is the duration between a post-ictal phase and a pre-ictal phase. Figure 4.1 displays the visual representation of random selected EEG signals from

two datasets used in this work. The screenshots are captured from a 30-second EEG recording, and the original signals are the voltage differences between each electrode pair listed on the left side of each image. The human can identify ictal segments by eye observation. For example, spikes before the seizure onset, quick activities when the seizure begins, and rapid amplitude changes. In current epilepsy studies, The long-term EEG monitoring needs to be manually annotated or curated by certified neurologists, which is both labor-consuming and time-consuming. An automatic, effective and precise seizure detection tool for long-term EEG recordings can help neurologists to quickly identify seizures. The tool may also change the procedure for annotating the EEG monitoring, a pre-annotated seizure list can be preliminarily generated before the EEG data being handed to neurologists. Considering the data explosion may happen to the EEG recordings, the automatic annotation tool can quickly transform the data into information to improve existing epilepsy research.

Another challenge for EEG-based seizure detection is that the EEG signal characteristics from each subject are unique like fingerprints. Training the model with features extracted across the cohort and testing on a new subject is a difficult task, which requires the features to be common measurements that can distinguish seizure and non-seizure but not different subjects. Moreover, most existing methods are segment-based, one limitation is that the EEG channels used for the model are fixed. If one or multiple of the channels contain significant noises or abnormal signals, or the EEG monitoring lacks certain channels, the data may not be applied to the model or the detection quality of the algorithm may be reduced.

In this chapter, we developed an automatic channel-based cross-patient seizure detection model using two long-term EEG signal dataset: Children’s Hospital Boston-Massachusetts Institute of Technology (CHB-MIT) with 23 patients and random selected 23 patients from the Center for SUDEP Research (CSR) [26]. Our contributions are the following:

- We built an automatic channel-based seizure detection model using 8-second segmentation and 135 features extracted from 961 hours (CHB-MIT) and 2093 hours (CSR) EEG signal data.
- We performed a cross-patient training strategy and three evaluation methods: channel-based, segment-based, and case-based. The model is tested on the 195 seizure cases and 98.73% of the non-seizure period, which is significantly larger than the testing of existing work.
- Our seizure detection model improved the seizure detection rate from 85% to 90.75% on the CHB-MIT dataset comparing to previous cross-patient model and achieved 92.23% overall accuracy, 93.57% specificity, 81.08% sensitivity, and 85.41% AUC on the segment-based evaluation.

4.2 Datasets

The first dataset is the Children’s Hospital Boston-Massachusetts Institute of Technology (CHB-MIT) Scalp EEG Database [21]. The recording covered 182 seizures in 192 files from 23 pediatric subjects (5 males, ages 3-22; and 17 females, ages 1.5-19). The sample rate is 256Hz with 16-bit resolution and all the data contains at least 23 EEG channels (24 or 26 in a few cases). For each subject, a summary file annotated start time and end time and all recorded seizures. CHB-MIT Scalp EEG Database is one of the most used EEG data for epileptic seizure detection research. In this study, we implemented our model on the CHB-MIT dataset and compared our performance with current work. The details of each patient from the CHB-MIT dataset are shown in the left part of the Table 6.1.

The second dataset we used in this work is from the CSR database, which contains patients of SUDEP or with a high risk of epilepsy death. The EDF data from CSR usually includes more than 60 channels of signals, such as EEG signals, EKG signals,

channels used in the CHB-MIT dataset. Each patient contains at least one annotated seizure. The details of each patient from CSR dataset are shown in the right part of Table 6.1

Table 4.1: The details of collected data of 23 patients from the CHB-MIT dataset and 23 patients from the CSR dataset. SZ = number of seizures, LSZ = number of lead seizures.

Subject	SZ	LSZ	Age	Duration	Subject	SZ	LSZ	Age	Duration
CHB-1	7	5	11	40:33:8	CSR-1	8	6	29	112:27:36
CHB-2	3	3	11	35:15:59	CSR-2	3	3	76	66:7:5
CHB-3	7	6	14	38:0:6	CSR-3	5	2	31	94:51:1
CHB-4	4	3	22	156:3:54	CSR-4	4	4	28	65:41:11
CHB-5	5	4	7	39:0:10	CSR-5	10	10	66	139:37:48
CHB-6	10	10	1.5	66:44:6	CSR-6	4	4	28	110:21:12
CHB-7	3	3	14.5	67:3:8	CSR-7	6	3	24	120:28:1
CHB-8	5	5	3.5	20:0:23	CSR-8	4	4	34	49:2:9
CHB-9	4	4	10	67:52:18	CSR-9	4	4	31	71:1:14
CHB-10	7	7	3	50:1:24	CSR-10	4	4	32	74:19:8
CHB-11	3	3	12	34:47:37	CSR-11	2	2	x	77:27:25
CHB-12	40	10	2	23:41:40	CSR-12	5	5	72	64:28:15
CHB-13	12	7	3	33:0:0	CSR-13	7	2	29	166:54:32
CHB-14	8	5	9	26:0:0	CSR-14	3	2	43	44:29:15
CHB-15	20	14	16	40:0:36	CSR-15	3	3	62	41:56:0
CHB-16	10	3	7	19:0:0	CSR-16	6	4	36	75:17:48
CHB-17	3	3	12	21:0:24	CSR-17	5	5	24	132:16:0
CHB-18	6	5	18	35:38:5	CSR-18	4	2	29	120:50:2
CHB-19	3	3	19	29:55:46	CSR-19	1	1	50	118:3:23
CHB-20	8	4	6	27:36:6	CSR-20	1	1	38	99:48:2
CHB-21	4	4	13	32:49:49	CSR-21	1	1	29	70:31:48
CHB-22	3	3	9	31:0:11	CSR-22	3	3	25	117:21:31
CHB-23	7	5	6	26:33:30	CSR-23	1	1	49	56:14:14
CHB-Total	182	119	N/A	961:38:20	CSR-Total	138	76	N/A	2097:34:12

4.3 Pre-processing

Our work only focused on classification on two classes of data: ictal segments and inter-ictal segments. For this work, we define:

- A segment is an 8-second window of EEG data containing multiple channel signals;
- An ictal segment is a window containing at least 4 seconds data inside of lead seizure onset and end duration;
- An inter-ictal segment is a window containing no data inside of seizure onset and end duration.

Here, a lead seizure is the first seizure after a 1-hour non-seizure period. If two seizures are close to each other, less than 1 hour as we defined, we drop the second seizure data. By applying this rule, we selected 119 out of 182 seizures from the CHB-MIT dataset, and 76 out of 94 seizures from the CSR dataset. The number of lead seizures for each subject is shown in Table 6.1.

The epilepsy seizure data from both of the CHB-MIT dataset and CSR dataset includes two parts. One is the digital signal data, which is stored as EDF files, while the other is annotation data, which is stored as text files. An EDF file consists of an EDF header that stores metadata for the EEG signals, following the digital data for each channel; the other one is an annotation file contains rows of annotations in the corresponding EDF file. For the CHB-MIT dataset, seizures are clear annotated one by one with start and end time in the corresponding files. For the CSR dataset, each row of an annotation file includes a timestamp and comprehensive annotation texts which are not limited to seizure onset and seizure end. Part of the text data is curated annotations using a standard annotation terminology. We only extract the terms of seizure start and end to locate the period of ictal segments.

Finally, using the seizure temporal information, we extracted continuous ictal and inter-ictal data as described, and split them into 8-second segments using a sliding window. To balance the number of ictal segments and inter-ictal segments during training, we use 3 sliding window steps: 1 seconds for ictal segments; 8 seconds for pre-ictal(1 hour before seizure onset) and post-ictal(1 hour after seizure end); and 30 seconds for other inter-ictal segments that are at least 1 hour away from ictal.

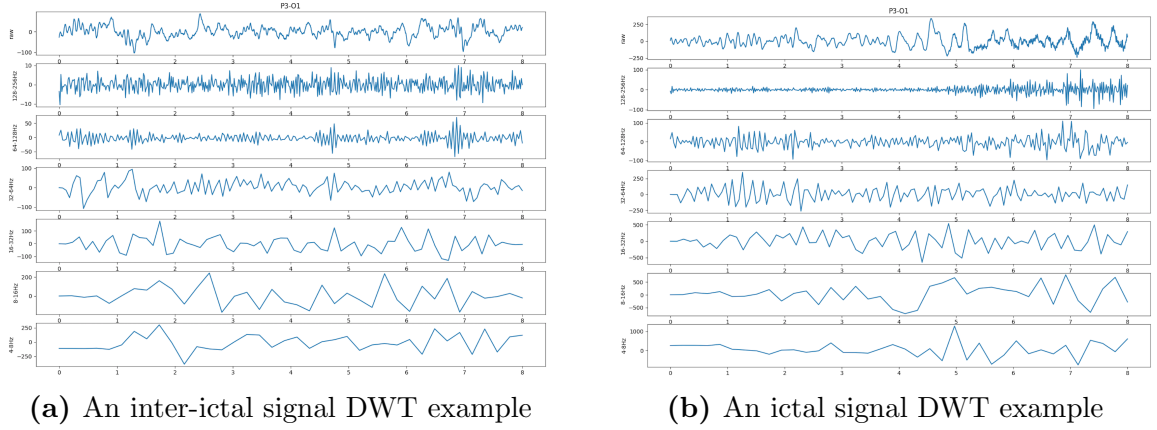


Figure 4.2: EEG signal DWT examples of 8-second EEG signal segment from channel P3-O1 from subject CHB-1. In both (a) and (b), the first row is the original 256Hz raw signal. From the second row to last row are sub-band coefficients: cD1(128Hz-256Hz), cD2(64Hz-128Hz), cD3(32Hz-64Hz), cD4(16Hz-32Hz), cD5(8Hz-16Hz), and cA5(1Hz-8Hz).

Most EDF files from CHB-MIT dataset use a montage with 20 channels: ‘FP1-F7’, ‘F7-T7’, ‘T7-P7’, ‘P7-O1’, ‘FP1-F3’, ‘F3-C3’, ‘C3-P3’, ‘P3-O1’, ‘FP2-F4’, ‘F4-C4’, ‘C4-P4’, ‘P4-O2’, ‘FP2-F8’, ‘F8-T8’, ‘T8-P8’, ‘P8-O2’, ‘FZ-CZ’, ‘CZ-PZ’, ‘T7-FT9’, ‘FT10-T8’. The duplicate channels ‘T8-P8’, ‘P7-T7’ was removed from processing. By the reason of 60Hz electrical artifacts, we applied a band filter with frequency between 57Hz and 63Hz to the original signal data. For an 8-second EEG signal segment with 22 channels, we extracted 135 features from each channel as describe in the following:

Raw signal. Raw signal is the original data collected from a montage channel. Each segment contains 2048 data points. For raw signal, we calculated 15 features: 1) median; 2) mean; 3) mean-minimum; 4) the 5th percentile of value; 5) the 25th

percentile of value 6) the 75th percentile of value 7) the 95th percentile of value; 8) the entropy of the distribution for signal values with occurrence probability; 9) the number of signal zero crossings; 10) the number of signal mean value crossings; 11) standard deviation; 12) variance; 13) root mean square; 14) kurtosis; 15) skewness.

Discrete wavelet transform (DWT) coefficients. We applied 5-level DWT on the raw EEG signal. At every level, the input is decomposed into two sub-band with halved frequency range: approximation coefficients A for the lower frequency band and detailed coefficients D for the higher frequency band. The data point number for both sub-band coefficients are reduced to half. Figure 4.2 shows two examples of 5-level decomposition on ictal signal and non-ictal signal using DWT. We calculated 15 features in every sub-band which are the same as the 15 features for the raw signal.

Frequency domain features. We transformed the raw signal into two types of frequency domain data. The first one is Power Spectral Density (PSD), which describes the distribution of spectral energy on frequencies. The second one is Fast Fourier Transform (FFT), which converts signal amplitude in the time domain to amplitude in the frequency domain. For the two frequency domain data series, we computed the top five peaks of energy/frequency and amplitude/frequency pairs as features.

Autocorrelation. The last transforming process is calculating the correlation of the raw signal with a time delay version of itself. The high autocorrelation shows the signal is similar to its copy after a gap of time. We computed all correlation/delay time pairs of the raw signal and extracted the first five correlation peaks as our features.

Table 4.2: CHB-MIT dataset testing results.

	Channel-based				Segment-based				Case-based		
Subject	acc	spe	sen	AUC	acc	spe	sen	AUC	acc	spe	sen
CHB-1	95.37	97.66	79.74	97.18	97.95	98.4	94.88	96.63	92.42	91.52	100
CHB-2	97.81	99.05	80.36	98.71	99.31	99.43	97.71	98.57	95.45	95.12	100
CHB-3	95.05	97.32	81.23	97.19	96.98	96.84	97.8	97.32	83.92	81.63	100
CHB-4	97.37	98.88	58.95	96.04	95.8	97.9	85.26	91.58	94.82	94.33	100
CHB-5	92.9	97.33	70.77	94.13	94.85	95.99	89.17	92.57	77.58	75.47	100
CHB-6	97.89	99.17	80.78	98.00	95.48	97.37	57.42	67.39	84.61	88.09	70
CHB-7	97.87	99.39	77.49	96.21	98.51	99.16	89.94	94.54	96.66	96.29	100
CHB-8	78.87	81.30	74.31	86.79	85.67	80.45	95.45	87.95	29.72	18.75	100
CHB-9	87.2	86.91	91.83	96.44	84.6	83.85	96.79	90.31	63.63	58.62	100
CHB-10	81.71	82.48	75.44	89.04	95.28	96.05	88.99	92.52	85.11	82.5	100
CHB-11	86.8	93.87	67.36	91.54	79.31	72.41	98.27	85.34	26.82	21.05	100
CHB-12	84.38	86.45	79.11	86.55	87.59	92.07	77.99	80.02	88.28	88.12	90
CHB-13	68.9	76.43	37.17	57.61	81.52	84.81	51.29	59.15	74.19	75	74
CHB-14	69.39	73.34	22.37	46.38	67.54	71.6	19.21	45.40	35.56	30	80
CHB-15	70.37	94.09	36.71	73.73	86.47	99.35	54.24	66.79	92.55	94.59	85
CHB-16	90.72	93.87	32.35	78.58	94.19	97.29	57.23	67.26	83.33	94.73	66.67
CHB-17	90.04	92.64	77.34	94.28	95.44	95.89	93.24	94.56	93.10	92.30	100
CHB-18	92.93	99.15	44.72	86.34	96.14	99.25	71.83	85.53	93.75	93.02	100
CHB-19	96.3	98.4	78.83	98.04	97.91	98.95	89.12	94.03	91.67	90.91	100
CHB-20	88.84	95.46	45.02	70.35	94.91	98.4	71.85	85.12	96	95.23	100
CHB-21	92.85	95.99	58.65	91.45	97.93	98.55	91.13	94.84	88.63	87.5	100
CHB-22	97.19	98.65	82.1	98.09	98.93	99.16	96.62	97.88	95	94.59	100
CHB-23	94.52	97.85	78.69	95.38	98.91	98.83	99.3	99.06	89.65	86.36	100
All	88.92	92.85	65.71	87.74	92.23	93.57	81.08	85.41	81.63	80.36	90.75

4.4 EEG Signal Classification

In this work, we applied eXtreme Gradient Boosting (XGBoost) [65] supervised machine learning technique for ictal and non-ictal classification. It provides a highly efficient implementation of gradient boosting framework: predicting by using an ensemble of multiple classifications and regression trees. The model uses features sub-sampling and adding regulation in the cost function to avoid overfitting. An approximate algorithm for split find is used for speeding up with lower memory usage. It also supports parallel computing to reduce training time costs.

In our processed channel-based dataset, one sample is a feature array of size 135 which is calculated from an 8-second signal segment of an EEG channel. Every sample is labeled by “ictal” and “non-ictal”. We first used a grid search for the hyperparameters fine-tuning on a sub-dataset, then trained the model with the whole dataset. The total collected numbers of ictal and non-ictal samples in the CHB-MIT dataset are 11,282 and 70,124. The total collected numbers of ictal and non-ictal samples in the CSR dataset are 6,711 and 177,533.

To evaluate the performance of each model, we used leave-one-out validation. For each testing subject, we train the model using all segment data from other subjects. We implemented three types of testing on the model. The first one is segment-based testing, which is a commonly used testing method in previous studies. Every processed segment from the testing subject is applied to the seizure detection model. According to model’s prediction, we calculated four metrics: 1) Accuracy = the number of correct segment prediction/ number of total testing segments; 2) Specificity = the number of correct non-ictal segment prediction/ number of total non-ictal testing segments; 3) Sensitivity = the number of correct ictal segment prediction/ number of total ictal testing segments; 4) Area under the curve (AUC) = the area percentage under the Receiver Operating Characteristics(ROC) curve with true positive rate

against the false-positive rate using threshold from 0 to 1. The next one is the channel-based testing. Since our model makes a weak prediction on each channel, the same four metrics can be calculated for all channel results.

The last testing is based on EEG recording cases. A full EEG recording from a patient can be split into four categories: 1) Ictal cases: from 4 seconds before seizure onset to 4 seconds after seizure end; 2) Pre-ictal cases: from 1 hour before seizure onset to 2 minutes before seizure onset; 3) Post-ictal cases: from 2 minutes after seizure end to 1 hour after seizure end; 4) all periods other than ictal, pre-ictal and post-ictal are inter-ictal cases. A single case may be naturally divided into multiple cases because there are gaps between EDF files. An 8-second sliding window with a 1-second sliding step is implemented during case testing. An ictal case is evaluated as pass if any 8-second segment in the case is predicted as positive by the model. A non-ictal case (inter-ictal, pre-ictal, and post-ictal) is counted as pass if none of all 8-second segments in the case is predicted as positive by the model. The case-based testing is an upper view of model detection performance on long-term EEG monitoring. We calculated three metrics for case-based testing: 1) Accuracy = the number of passed cases/ number of total testing cases; 2) Specificity = the number of passed non-ictal cases/ number of total non-ictal cases; 3) Sensitivity = the number of passed ictal cases/ number of total ictal cases.

Accuracy is used to evaluate the overall correct labeling rate of the model. Specificity shows the performance of the model when recognizing non-ictal data. Higher specificity implies a lower false detection rate. Sensitivity expresses the accuracy of the model when facing ictal data. Higher sensitivity implies a low rate of missing detection. The AUC indicates the ability of the model to distinguish the ictal and non-ictal class.

4.5 Result

The aim of this study is to build a channel-based classifier of ictal and non-ictal EEG signals that can be applied to long-term seizure detection on EEG monitoring. The model was evaluated using the CHB-MIT dataset and CSR dataset on three types. The channel-based and segment-based evaluation used all ictal, pre-ictal and post-ictal data, and part of inter-ictal data (extracted every 30 seconds). The case-based evaluation used all recorded EEG data except the 2-min gap between pre-ictal and ictal, and between ictal and post-ictal.

The experimental results of each patient in the CHB-MIT dataset are listed in Table 4.2. The last row displays the overall measures of the database. Measures of channel-based and segment-based testing are the average of every subject’s results, and the measurements of case-based are calculated by the accumulative case count of all the subjects. Because the number of ictal labels and non-ictal labels are unbalanced, the values of accuracy and specificity are very close. The segment-based model is the ensemble of a channel-based model, so the overall segment-based performance is improved from channel-based. The accuracy and specificity show consistency in almost every subject. 18 of 23 subjects results in specificity above 90% on segment-based testing. however, the sensitivity varies with the subject because of different seizure characteristics. While subjects 1,2,3,8,9,11,17,21,22,23 were observed with sensitivity above 90%, subject 14 only has 19.21% sensitivity. Case-based testing shows the model’s performance on continuous EEG recording. The case-based specificity is extremely sensitive to false positives, one false detection can fail the whole non-ictal testing case. Meanwhile, the testing allows missing detection of ictal segments, which leads to an increment of sensitivity comparing to segment-based testing. 17 Subjects detected 100% seizure during the long-term testing and 10 subjects obtained $\geq 90\%$ overall accuracy.

Table 4.3: CSR dataset testing results.

	Channel-based				Segment-based				Case-based		
Subject	acc	spe	sen	AUC	acc	spe	sen	AUC	acc	spe	sen
CSR-1	94.78	97.2	73.49	93.58	97.56	98.99	84.79	91.89	89.36	89.73	87.5
CSR-2	95.85	97.85	37.31	84.39	97.61	99.49	41.78	70.64	86.96	85	100
CSR-3	96.96	98.42	20.89	81.24	98.27	99.78	19.51	59.65	86.67	92	60
CSR-4	96.32	97.87	63.01	90.28	98.22	99.45	71.66	85.55	92.59	95.65	75
CSR-5	96.5	98.07	65.31	90.54	97.89	99.07	74.09	86.58	88.14	85.71	100
CSR-6	98.28	99.34	52.22	94.34	99.13	99.73	71.89	85.81	97.06	96.67	100
CSR-7	91.64	95.65	31.7	73.14	93.97	98.89	19.81	59.35	77.5	79.41	66.67
CSR-8	98.94	99.81	76.51	98.49	99.4	99.98	84.26	92.12	100	100	100
CSR-9	88.43	88.92	38.92	77.72	95.09	95.51	50.7	73.11	68.97	68	75
CSR-10	97.75	98.5	38.28	94.02	98.74	99.41	45.26	72.33	90.91	89.66	100
CSR-11	98.62	99.48	57.49	95.44	98.8	99.52	64.62	82.07	82.61	80.95	100
CSR-12	94.91	98.26	45.07	85.60	96.62	99.49	53.23	76.36	87.5	85.18	100
CSR-13	87.01	88.72	24.82	70.25	88.37	90.14	34.24	56.83	66.67	63.83	85.71
CSR-14	93.94	94.18	23.68	83.37	93.26	93.4	42.5	82.94	57.89	50	100
CSR-15	92.07	97.91	3.17	68.36	93.78	99.69	3.24	51.46	85.71	96.55	33.33
CSR-16	97.7	98.85	62.33	91.51	98.9	99.96	66.39	83.17	100	100	100
CSR-17	96.09	97.34	14.01	76.90	97.38	98.29	29.23	63.76	78.94	75	100
CSR-18	96.5	97.04	13.51	77.40	96.34	96.69	29.41	63.05	75	78.57	50
CSR-19	98.36	99.44	44.44	86.74	98.79	99.78	48.82	74.30	95	94.73	100
CSR-20	99.2	99.89	54.97	89.37	99.35	99.94	61.84	80.89	100	100	100
CSR-21	79.92	80.26	71.62	86.45	84.52	84.74	83.78	84.28	85.71	83.33	100
CSR-22	97.46	99.46	50.68	92.65	98.68	99.85	91.89	85.46	96	95.45	100
CSR-23	98.87	99.91	66.81	95.92	99.24	99.96	76.92	88.44	100	100	100
All	95.05	96.62	44.79	85.99	96.52	97.90	54.34	76.09	85.69	85.78	85.53

Table 4.3 reports the model performance results of each patient in the CSR dataset. Comparing to the CHB-MIT dataset, the results show a more varied detection quality due to multiple data sources of the CSR dataset. Such a reason results in low average sensitivity of 54.34% for segment-based testing. On the contrary, good results still obtained on specificity (averaging 96.62% and 97.90% for channel-based and segment-based). The robust performance on non-ictal data is proved in the case-based evaluation: during the 2093 hours testing, the pass rate of non-ictal cases is 85.78%. Perfect results were obtained for subjects 8, 16, 20, and 23 (about 280 hours total recording length), which means all seizures were detected and no false alarm occurred.

4.6 Discussion

In this work, we developed an automatic channel-based cross-patient seizure detection model on EEG signals. The model has two potential usages from clinical aspect, one is a embedded component in EMU for real-time seizure alarming, another usage is automatic marking of seizures on unannotated EEG signal files to fasten the manual annotation process. Our work provides a new evaluation methods and results using continuous long-term EEG recordings in CSR dataset, which can be a benchmark for future seizure detection study using EEG signals.

We compared the performance on seizure detection of our method with the results of other existing methods. Table 4.3 lists the comparison result, all measures are the results of the segment-based evaluation. Nine papers using the CHN-MIT database were included and their methods are briefly described in the background section. Since the only other cross-patient method used one long segment for each seizure (segment = case under the circumstance), our average case-based sensitivity 90.75% is higher than theirs. Comparing with inter-patient and patient-specific models, even our task is more challenging, our model’s performance still ranks in the middle. From

another aspect, our model was tested on almost the whole recording from the dataset, which significantly overcomes other testing data usage ratio from 0.4% to 5.87%. Our evaluation based on long-term data simulates the model performance on real-world EEG monitoring.

Our cross-patient model shows the seizure detection methods may result in varied performance on different. One limitation of our model is that our model did not consider the variety between subjects. Future work can be utilizing the demographic data of subjects to cluster subjects into different cohorts. Then improved model can be built based on the cohort. Besides, seizure types may be an important feature to affect seizure detection quality. A more precise seizure type-specific model may have better performance.

Table 4.4: Comparison table for testing results on CHB-MIT dataset.

Study	acc	spe	sen	AUC	Model Type	Non-ictal Data Tested(%)
Kiranyaz et al. 2014 [50]	x	89.01	94.71	x	patient-specific	5.87
Fergus et al. 2015 [47]	x	88	88	93	inter-patient	1.99
Xun et al. 2015 [51]	77.07	x	x	88.8	inter-patient	2.19
Thodoro et al. 2016 [48]	x	x	85	x	cross-patient	0.4
Yuan et al. 2018 [52]	96.61	x	x	98.47	inter-patient	2.19
Zhou et al. 2018 [54]	97.5	96.9	98.1	x	inter-patient	x
Park et al. 2018 [49]	85.6	91.7	80.6	x	inter-patient	4.33
Alickovic et al. 2018 [12]	100	100	100	x	inter-patient	3.1
Tian et al. 2019 [53]	98.3	96.7	99.1	x	patient-specific	x
This work	92.23	93.57	81.08	85.41	cross-patient	98.73

4.7 Conclusion

In this chapter, we propose a channel-based model for automatic cross-patient seizure detection based on the EEG signal. We trained and tested our model using two scalp EEG datasets: 1) a widely used public epilepsy dataset CHB-MIT dataset containing 119 lead seizures and 961 hours EEG data; 2) a diverse cross-site epilepsy dataset

CSR dataset including 76 lead seizures and 2093 hours EEG recording. We extracted 135 features from each 8-second channel signals using signal processing methods like DWT, autocorrelation, PSD, and FFT and used such features with labels to train an XGBoost model. The model is evaluated using 195 seizure cases and 98.73% of the non-seizure period, which is significantly larger than the testing of existing work. The case-based testing results show that our model detected 90.75% lead seizures, and achieved 92.23% overall accuracy, 93.57% specificity, 81.08% sensitivity, and 85.41% AUC in the segment-based evaluation. The results show our cross-patient performance of accuracy and specificity is among the top patient-specific and cross-patient models from the currently proposed works.

CHAPTER 5. Seizure Prediction on Scalp EEG Data

5.1 Motivation

Epileptic seizures have been marked as an “unpredictable” disorder because no practical tool was available to reliably predict seizure onset in real time [55].

For decades, long-term brain activities of epilepsy patients have been systematically captured by monitoring patients’ electrophysiological signals including Electroencephalography (EEG). EEG records voltage fluctuations resulting from ionic current within the neurons of the brain [3]. Recent approaches use time-frequency analysis for epilepsy studies. Time-frequency approach transforms EEG data into spectrogram, a heat map of the spectrum of signal frequencies varying with time. Figure 5.1 displays the visual representation of EEG spectrogram images from 15 bipolar montage channels. The images are generated from a 30-second EEG recording, and the original signals are the voltage differences between each electrode pair. The human eyes can identify ictal segments by finding areas of high power with high frequency, but it remains hard to distinguish pre-ictal from inter-ictal segments. How to effectively and precisely detect pre-ictal period before epileptic seizures using EEG data is an important challenge.

Since the 1970s, researchers successfully extracted relevant features to recognize a seizure from EEG recordings [6, 7]. EEG signals are also analyzed in the frequency domain for classification [8, 9]. Additionally, traditional machine learning methods, such as support vector machine (SVM) and random forest, made progress on seizure classification and prediction [10, 11]. In recent years, deep learning gains popularity in seizure analysis with large-scale datasets. In 2014, the American Epilepsy Society, Epilepsy Foundation of America, National Institutes of Health and a data science competition platform kaggle.com (Kaggle, Inc. New York NY, USA) together launched a competition to predict seizure with 1-hour lead time using seizure data from five

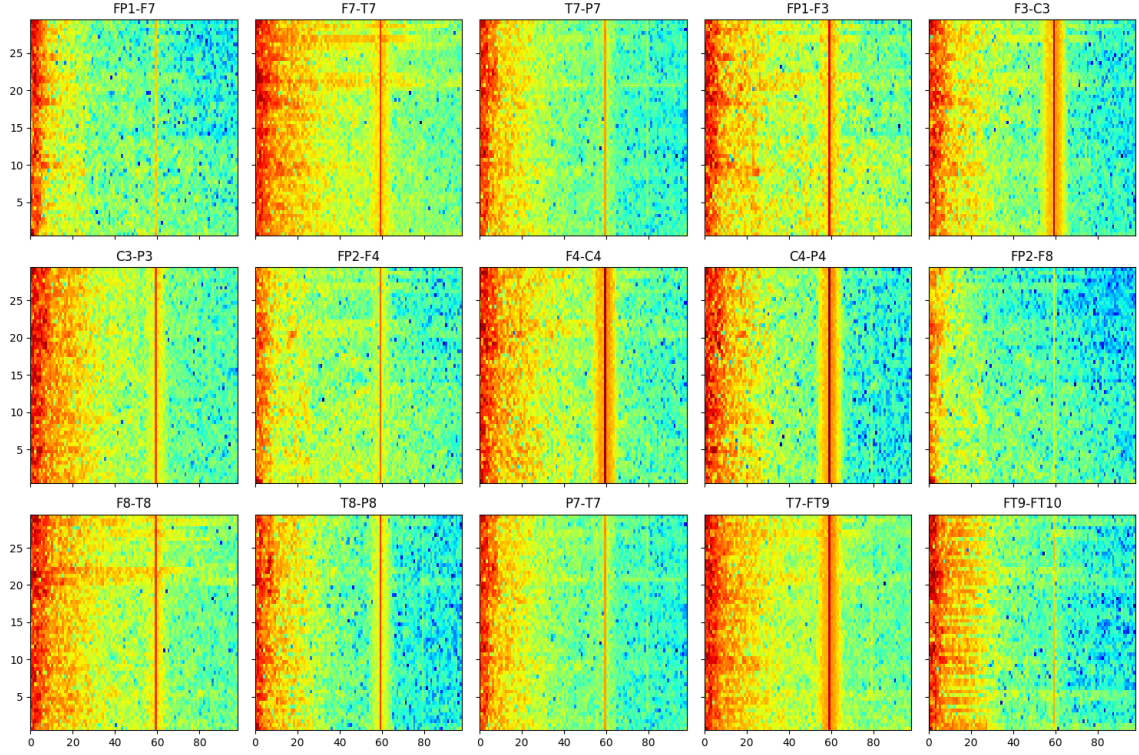


Figure 5.1: An example of 15-channel spectrogram images. For each image, the x-axis is frequency and the y-axis is time. A red point indicates higher energy at the time and frequency, and the blue means a lower energy point. At around 60Hz, a power line exists in most images. Such noise is eliminated during data pre-processing.

canines and two humans. Two artificial neural networks (ANN) methods appeared in the final top 10 [13]. Convolutional neural networks (CNN) with spectrogram is used to classify pre-ictal and non-ictal EEG segments and achieved high performance on patient-specific models [66]. In addition, wearable devices for warning seizures has been developed and tested [15, 16]. Mobile devices capturing seizure data offer great opportunities for researchers to implement seizure prediction algorithms.

However, current seizure prediction work has two limitations: 1) although many public seizure datasets consist of long periods of recording, the pre-ictal data is relatively small because the number of recorded seizure episodes is small; 2) multi-channel epilepsy data is larger than the data used in typical machine learning studies. The processing and training of such data are more complicated and time consuming.

In this work, we developed an efficient transfer learning model to extract and pre-process EEG data, train from a base model, and evaluate the trained model using 17 patients’ data from the Center for SUDEP Research (CSR) [26]. Our contributions are the following:

- We built transfer learning prediction models using two pre-trained deep learning architectures and the training time is significant reduced;
- We performed a multi-channel input evaluation with a decision queue using channel voting;
- Our transfer learning models reached 86.79% sensitivity and 3.38% false-positive rate, which indicated the transfer learning setup achieved high prediction performance using patient-specific data.

5.2 Datasets

The dataset we used in the work is from the CSR database, which contains 408 subjects with 1,622 annotated seizures. The CSR data includes more than 60 channels of signals, such as EEG signals, EKG signals, blood pressure, peripheral capillary oxygen saturation (SpO2), etc. Because the EEG data in CSR are collected from different sites, the recording configurations (for example, sample rate) are not the same. We only used the data from the largest branch in the CSR dataset, University Hospital of Cleveland. Our work only focused on classification of two classes of data: pre-ictal segments and inter-ictal segments. For this work, we define:

- Pre-ictal segments are EEG recordings in the one-hour period before a lead seizure onset with a five minutes gap;
- Inter-ictal segments are selected from the period at two hours after a seizure onset and two hours before a seizure onset.

Here, a lead seizure is the first seizure after a 1.5-hour non-seizure period. An example of pre-ictal segments and inter-ictal segments extraction is illustrated in Figure 5.2. To better distinguish pre-ictal data from ictal signals, we added a 5 minutes gap between a pre-ictal segment and the seizure onset point. In the example, we did not extract the pre-ictal segment during the period between Seizure Onset 2 and Seizure Onset 3 because the gap between the two seizure is less than 1.5 hours. We did not extract the inter-ictal segment during the period between Seizure Onset 1 and Seizure Onset 2 because the gap between the two seizure is less than 4 hours. By the reason that part of the patients' data is split into pieces by non-recording period, we did not include seizures that occurred in the first hour of a continuous recording. We randomly select 20 subjects' data from the database, each one contains at least three seizures and recorded with a sample rate of 200Hz. According to our lead seizure definition, 3 patients only have 1 pre-ictal segment, so we removed them from our final results.

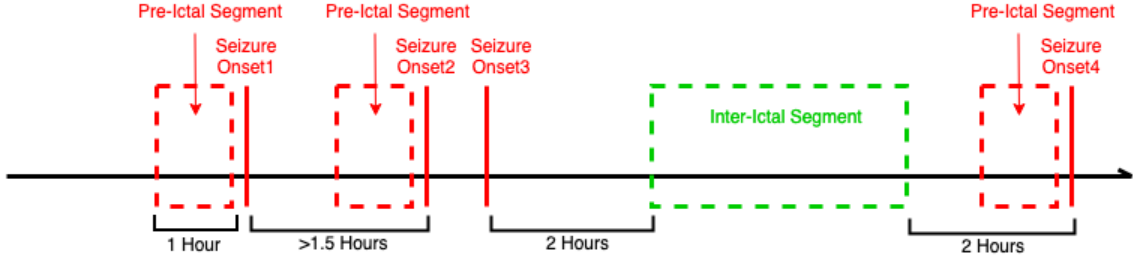


Figure 5.2: An example of our pre-ictal segments and inter-ictal segments extraction method. The horizontal black line is the timeline. The vertical red lines indicate the start points of seizure. The red dashed boxes cover the periods of the pre-ictal segments, and the green dashed box covers the period of the inter-ictal segment.

Data statistics of the 17 patients from CSR dataset are shown in Table 6.1. The total pre-ictal segments used in the experiment are 53. The average EEG recording length from the 17 patients is above 88 hours. The epilepsy seizure data from the CSR dataset is stored in two parts: digital signal data is stored as EDF files while

Table 5.1: The summary of the 17 patients randomly selected from the CSR dataset.

Patient	Seizure number	Lead seizure number	Total recording time (Hours)
Patient1	4	4	74.29
Patient2	4	4	77.43
Patient3	7	2	166.9
Patient4	5	2	112.44
Patient5	3	3	66.11
Patient6	14	2	44.48
Patient7	4	3	132
Patient8	5	2	94.85
Patient9	4	3	65.68
Patient10	8	5	139.62
Patient11	5	4	110.34
Patient12	3	3	41.93
Patient13	6	2	120.47
Patient14	4	3	49.03
Patient15	6	4	64.45
Patient16	4	4	71.02
Patient17	6	3	75.28
Total	92	53	1506.32

annotation data is stored as text files. An EDF file consists of an EDF header which stores metadata for the EEG signals, following the digital data for each channel. An annotation file contains rows of annotations in the corresponding EDF file, each row includes a timestamp and an annotation term. All the files for a visit are compressed into a zip file. In CSR dataset, the annotation files stored comprehensive text data, which includes curated annotations using a standard annotation terminology. To generate our training and testing datasets for this work, we first read all annotation files, then loaded and stored all time stamp of “seizure onset” and “seizure end” annotations. Next, we used the timestamps to locate the seizures from start to stop,

which is the period of ictal. Finally, using the seizure temporal information, we extracted pre-ictal and inter-ictal data as described, and split them into 30-second sliding window clips. To evaluate the performance of each model, we used k fold validation, where k is the number of lead seizures for each patient. Each testing dataset includes a 1-hour pre-ictal segment and a 5-hour inter-ictal segment. The other pre-ictal segments and inter-ictal segments are used for the training process. Because the total time of inter-ictal segments is always large than the total time of pre-ictal segments, we used 15 seconds as sliding step length to enlarge the size of training pre-ictal data and use no overlap for inter-ictal data. 80% of training pre-ictal clips and the same number of training inter-ictal clips for each patient are fed to train the patient-specific models, and 80% of training data are used for model validation during training. As a result, query, unzipping, and data extraction becomes time-consuming procedures. To build a more researcher-friendly seizure data platform, we developed the SeizureBank as described.

5.3 Pre-processing

The CSR data uses referential montages that the channels share a reference electrode. We first selected and translated the CSR referential montages to 15-channel bipolar montages, and then split them into 5 groups: a) “FP1-F7”, “F7-T7”, “T7-P7”; b) “FP1-F3”, “F3-C3”, “C3-P3”; c) “FP2-F4”, “F4-C4”, “C4-P4”; d) “FP2-F8”, “F8-T8”, “T8-P8”; e) “Fp1-FZ”, “FZ-CZ”, “CZ-PZ”. Each group is an input channel and the output is a channel vote. Then we used Short-time Fourier Transform (STFT) to generate the time-frequency domain data after splitting data into 30-second clips for each channel in each group. Because the sampling rate of our dataset is 200Hz, the spectrogram image only covers the frequency from 1 to 100Hz. Also, we removed the frequency between 57Hz and 63Hz because the common high power noise is around 60Hz.

5.4 Classification Model

We developed a transfer learning model to classify 30 seconds EEG spectrogram images into two categories: Pre-ictal or Inter-ictal. The source data is a time-series multi-channel EEG recording. We processed (including data crop, frequency filtering, and the short-time Fourier transform) the data into multi-channel spectrogram images. For the transfer learning model, we choose VGG19 and ResNeXt50 as the base model to extract the spectrogram image features, and added fully connected layers on the top of the base model, then output a prediction of pre-ictal or inter-ictal classification. After our transfer learning model is trained, we implemented a decision queue as a status checking window to make a final decision: warning or safe, which is a more intuitive result for a continuous testing stream. With the benefits of transfer learning and the decision queue, the system can make a continuous seizures warning during a pre-ictal period, and indicate safe status during an inter-ictal period.

Data preprocessing creates five 3-channel spectrogram images for each 30-second EEG data. Every image is labeled with pre-ictal or inter-ictal. For a patient with k lead seizures, we first selected the first pre-ictal segment and randomly select a 5-hour continuous inter-ictal segment as the testing data, then use other $(k-1)$ pre-ictal segments and rest inter-ictal segments as the training. To balance the training data, we randomly selected a set of non-continuous inter-ictal data so that the number of the two classes is the same. We repeated the training process for k times, so k testing models were built for the patient with k lead seizures. The pre-trained model weights using ImageNet dataset were loaded and froze at the start of the training process. VGG19 and ResNeXt50 are the two pre-trained models we used in the experiments. The top of the transferred model was replaced with 3 customized fully connected layers and output the probability of each category. Every model is trained for 25 epochs, and we selected the weight with the highest training accuracy for evaluation.

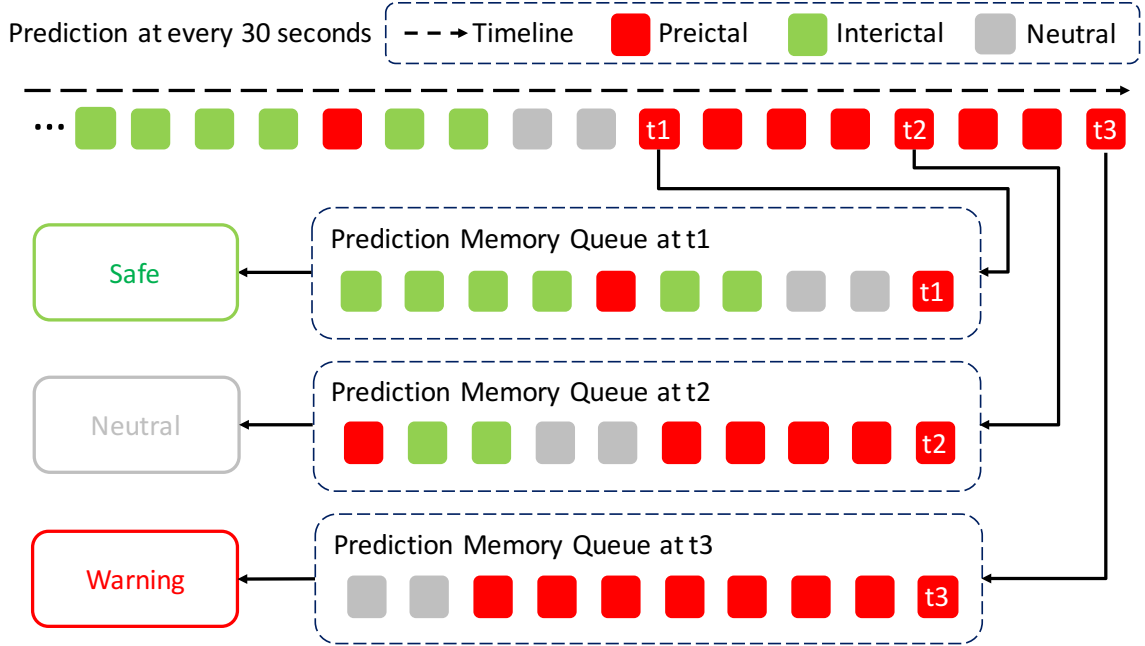


Figure 5.3: An example of prediction decision queue with threshold of 8. At the top is the time-series prediction results of input clips using pre-ictal detection model. The dotted box displays the status of the prediction decision queue at t_1, t_2 and t_3 , and the solid line boxes on the left show the final prediction results at t_1, t_2 and t_3 .

Instead of using general 1-input and 1-output model, our model used a channel voting strategy with 5-input and 5-output. For each given 30 seconds EEG data, we can generate 15 spectrogram images using EEG montage, and then split them into 5 input channels, each channel contains 3 images. Next, our model can output 5 votes for each input channel. If more than two of the outputs vote to pre-ictal, we mark it as red for the prediction; if the pre-ictal vote number equals to 2, we mark it as white; if all 5 votes are inter-ictal, we mark it as green. Then the output will be a series of flags with red, white, and green. If we decide the final results according to every single flag, a wrong pre-ictal prediction (false positive) will lead to a wrong seizure warning even if the wrong prediction occurs only once. To reduce the chance of reporting wrong pre-ictal warning by our model, we added a decision phase after the training process. As shown in Figure 5.3, the decision phase produced a series of flags. We used a prediction decision queue to make a final decision. The queue

is constructed with a small memory that stores recent vote results of pre-ictal and inter-ictal. The total number of red flags in the queue is calculated and a warning will be triggered when the number is above a threshold.

Figure 5.3 displays an example of how the final prediction is made by the prediction decision queue. In our experiments, the size of the decision queue is 10 and threshold of red flags in the queue is 8. At time t_1 , two pre-ictal prediction (red) occurs, but six outputs are inter-ictal. Under this circumstance, the number of red flags is 2, so the final decision is safe. Similarly, at time t_2 , more red flags entered into the queue but the number was still under the threshold, so the final decision is neutral. Finally, at time t_3 , eight outputs in the queue are red and two are white, then the final output turns to a seizure warning. For our evaluation, in a pre-ictal testing segment, if a warning decision is made, then the coming seizure is predicted. On the contrary, an unsuccessful seizure warning is counted if no seizure occurs during the pre-ictal testing period.

We used labeled continuous EEG data from 17 patients in the CSR dataset for the testing. The total number of tested lead seizure is 53, and the total tested inter-ictal duration is 265 hours. To evaluate the performance of our seizure prediction methods, we calculated two commonly used metrics in the previous seizure prediction research. They are Sensitivity and False Positive Rate (FPR). Their formulas are listed in the following:

- $Sensitivity = \frac{Number\ of\ Warned\ Seizures}{Number\ of\ Total\ Seizures}$,
- $FPR = \frac{False\ Positive\ Number}{Total\ Prediction\ Number}$,

Sensitivity equals to the true positive rate of the testing. In our case, sensitivity is the ratio of the correct final warning decision and total seizure number. FPR equals to the ratio of the total number of incorrect warnings and the total possible decision numbers during testing inter-ictal segments. Based on the requirements we

mentioned, our expectation will be the results of high sensitivity and low FPR. The details of the evaluation results are shown in the next section.

Table 5.2: The performance of the transfer learning with two base models: VGG19, ResNeXt50. Two measurements are Sensitivity and FPR (False Positive Rate).

	Truong [66]		VGG19		ResNeXt50	
Subject	Sen (%)	FPR (%)	Sen (%)	FPR (%)	Sen (%)	FPR (%)
Patient1	75	4.44	100	2.43	100	1.54
Patient2	75	3.70	100	2.58	100	0.77
Patient3	50	1.41	100	6.25	100	0.0
Patient4	100	0	100	6.67	100	2.58
Patient5	100	0.82	100	0.0	100	1.03
Patient6	50	4.34	0.0	4.62	0.0	2.87
Patient7	66.67	1.0	66.67	0.0	33.33	0.0
Patient8	100	6.12	50	3.54	100	0.98
Patient9	66.67	1.22	66.67	0.0	100	3.89
Patient10	40	0.28	80	0.34	80	4.75
Patient11	75	2.97	75	5.37	100	3.71
Patient12	100	7.42	100	6.15	100	4.04
Patient13	100	7.07	50	0.14	100	0.17
Patient14	100	7.04	66.67	0.40	66.67	2.24
Patient15	100	5.71	100	4.64	100	4.95
Patient16	100	0.0	100	0.23	100	0.17
Patient17	100	2.78	100	4.06	100	6.43
Average	79.25	3.16	83.02	2.79	88.67	2.13

5.5 Result

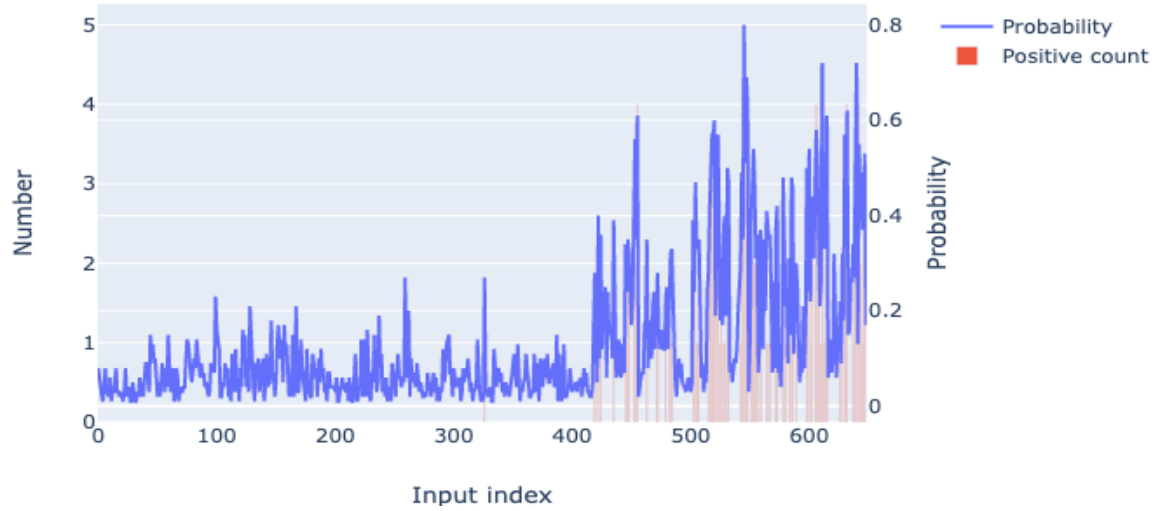
Our experiments have been developed using TensorFlow version 1.6. The testing environments are macOS High Sierra operating system, 2.7GHz Intel Core i5 CPU, and 8GB RAM. Testing speed can be improved by using higher-end hardware and GPUs in the future. From a total 20 random selected patients from CSR, we eliminated 3 of them with only one lead seizure. The results from 17 patients are shown in this section.

Seizure warning performance. Table 5.2 displays the evaluation results from selected 17 patient from the CSR dataset. Both models perform excellent on almost half of the patients (1, 3, 4, 5, 12, 15, 16, and 17), they predicted every seizure during the test. However, for patient 6, neither of the models caught the two testing seizures. In general, both VGG19 and ResNeXt50 model produced low FPR (3.4% and 3.38% on average), and ResNeXt50 model performed slightly better on sensitivity, which is 86.79% on average.

Figure 5.4 shows the real-time monitoring history of two testing examples. The x-axis is the index of the continuous 30-second sliding inputs, the y-axis is the value of two results' representation (probability and the number of positive votes from 5 input channels) at of each input. The probability (the blue line) is calculated by averaging the softmax output of positive (pre-ictal) from 5 channels, it ranges from 0-100%. The positive count (red bars) is the sum of the number of channels that output positive predictions. On the left end part of Figure 5.4a, even with a short period of low average positive probability, a correct warning is generated because of a list of continuing outputs with positive vote larger or equal to 2. In Figure 5.4b, we can see many spikes of high probability points, but no incorrect warning is generated because they are filtered by the decision queue.



(a) Pre-ictal testing



(b) Inter-ictal testing

Figure 5.4: Monitoring of the evaluation results of patient 2 using ResNeXt50 transfer learning model : (a) a 1-hour pre-ictal testing segment, (b) A 2.5-hour inter-ictal testing segment.

5.6 Discussion

Clinical significance. In this work, We introduced a transfer learning approach to predict seizures using EEG signals in the time-frequency domain. Since epilepsy is “unpredictable” and the complete reasons that trigger a seizure is still unclear. Our results show a possible solution by recognizing abnormal pre-ictal EEG signals. The

lead time of our seizure prediction allows patients and doctors to prepare for the coming seizures which could reduce the risk of injuries or death in real lives.

Definition of pre-ictal period. There is no formal definition of pre-ictal in the seizure prediction domain. Lead time before seizure can be from minutes to hours. How to determine a desirable lead time of seizure onset is an on going question. If the lead time is too short, the prediction problem is closer to a detection problem. If the time is too long, the biological mechanism may be uncoupled.

Training efficiency. One advantage of transfer learning is the significantly reduced training time if using the pre-trained parameter weights. For each testing model, we set the maximum training epoch number to 25. In other words, the model only learned the whole dataset 25 times, which is much less than the usual training process for the same architectures without pre-loaded weights.

Trigger of the seizure warning. In this work, we used an intuitive structure and prediction decision queue to control the final results after using the trained model. A more intelligent function can be developed in this part for a better result. However, even the best model may predict a false result. Because of the importance of the high sensitivity, ignoring a positive answer is always critical for seizure prediction.

Transfer learning fine-tuning. During the training process, we loaded the ImageNet pre-trained weight directly and only trained the top fully connected layers. By the reason that ImageNet classification and EEG spectrogram classification have differences, fine-tuning of the pre-trained weights is a potential way to improve the result.

Prediction quality. Another limitation is that our work only performed the prediction within the 1-hour pre-ictal period by given a “True or False” answer. Predicting seizure in one hour period provides a time delay for doctors and patients to use medical methods that can prevent seizure. However, a more precise prediction of the time before a seizure is more important and useful. For example, if a 60-minute lead time

seizure warning occurs when a patient is driving, the patient may choose to drive to a safe place, for example, home or hospital; on the other hand, if the warning is a 5-minute lead time seizure warning, the patient needs to stop the vehicle as soon as possible. In addition, a warning indicates a potential seizure “in 60 minutes” and “exactly after 60 minutes” is different. “In 60 minutes” means an uncertain period of time, while “exactly after 60 minutes” requires very high accuracy. In this project, we reduced the “exactly” problem to the prediction of a seizure within 5-minute lead time and provide a method to output a reliable prediction under one minute by partially analyzing a EEG segment. Next step is to precisely calculate the probability of a coming seizure, for instance, weather forecast, which is more reasonable in real-world cases. Moreover, estimating the exact time before the seizure onset is another future work.

5.7 Conclusion

In this chapter, we proposed a framework, including data pre-processing, a multi-channel input transfer learning approach using pre-trained models, and a prediction decision queue using multi-channel voting. We trained a patient-specific model with the deep learning approach from the CSR EEG database leveraging CSR’s large-scale labeled epilepsy patient data. A seizure prediction evaluation was performed using 53 pre-ictal EEG signal segments and 265 hours inter-ictal EEG data. Our results showed that after 25 epochs training, the transfer learning model using ResNeXt50 pre-trained model reached 86.79% sensitivity and 3.38% false-positive rate, which indicated the transfer learning setup for patient-specific seizure prediction is efficient and convincing.

CHAPTER 6. Seizure Localization on Stereoelectroencephalography Data

6.1 Motivation

The ultimate goal for the epilepsy study is to find the reason for seizures and prevent it from happening again. In the clinical site, surgeries include vagus nerve stimulation (VNS), focal cortical resection, lobectomy, hemispherectomy and corpus callosotomy [67], may be the part of the treatment. If the resected area contains the brain part that leads to initiating the seizures, the patient will be seizure-free at least for years. The epileptogenic zone by definition is “the area of cortex that is necessary and sufficient for initiating seizures and whose removal (or disconnection) is necessary for the complete abolition of seizures” [56]. How to fast and accurately locate the epileptogenic zone remains a big challenge for epilepsy patient treatment.

Nowadays, the epilepsy monitoring unit (EMU) provides evaluation and diagnosis to locate the seizure activities in the brain. By the reason that seizures demonstrate differently among types and individuals, sometimes determination of the resection zone becomes crucial and challenge. Increase the surgeries’ successful rate can reduce the pain for patients and the risk of brain damage. Stereoelectroencephalography (SEEG) is the gold standard for seizure localization in epilepsy studies. Neurologists and neurosurgeons can recognize seizures by looking at the time domain SEEG signals. Figure 6.1 represents SEEG signals from two different channels at the same time. The neurologists annotated the start time (red dotted line in the middle) of the seizure according to the seizure activity occurs in the red color channel. After diagnosis and the surgery, if the patient became seizure-free, we can infer the epileptogenic zone is inside the resection zone. The data of such seizure-free patients are valuable for prospective studies related to seizure localization.

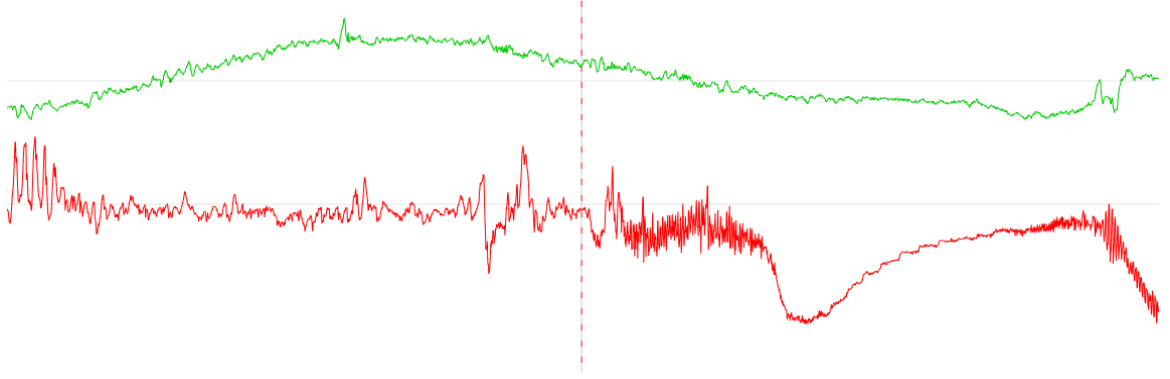
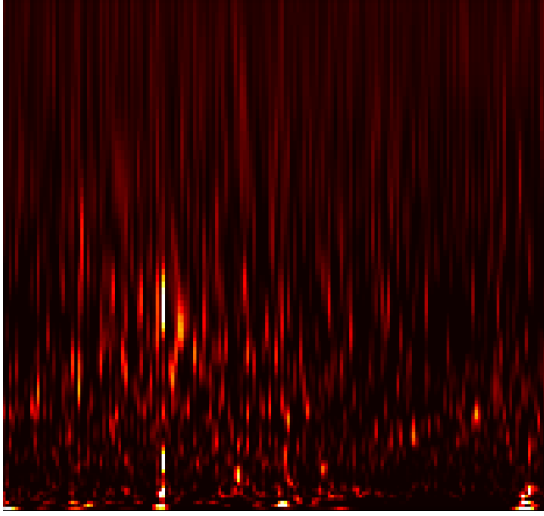


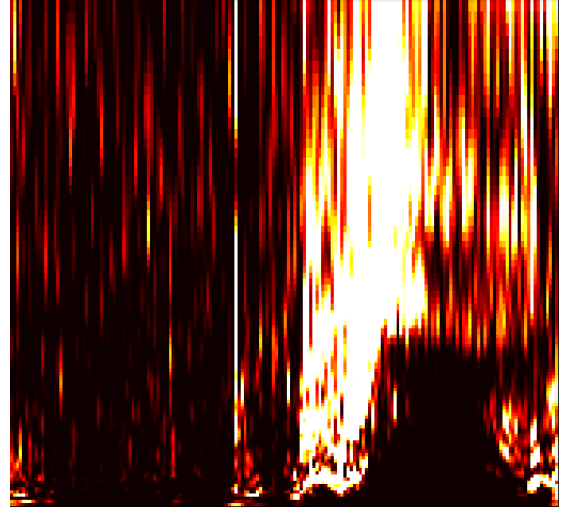
Figure 6.1: Time domain SEEG signals of two channels in a 40-second window. The channel in green is outside the epileptogenic zone and the channel in red is outside the epileptogenic zone. The red dotted line in the middle denotes the start point of the seizure.

Many studies are exploring convincing epileptogenic zone localization methods by state-of-the-art SEEG processing technologies and algorithms. Grinenko et al. Figure [14] developed a pipeline to locate the epileptogenic zone using SEEG signals. The authors discovered a specific ictal pattern of channels with seizure activities in the time-frequency domain and called it a fingerprint of the epileptogenic zone. The pattern includes three characteristics: 1) sharp transients or spikes; 2) multi-band quick activity concurrent; 3) suppression of lower frequencies. Figure 6.2 illustrates two time-frequency images. Figure 6.2b displays a fingerprint example in the dataset of our work. As a comparison, Figure 6.2a illustrates an example of a non-epileptogenic zone signal without the three characteristics. To extract such features, they applied the Morlet wavelet transform to SEEG data near seizure onset. After filtering, ridge detection, and masking, they extracted or computed frequency, timing, and areas to describe the processed data. Finally, an SVM classifier was trained using a dataset consists of 17 patients' SEEG data. The results show the fingerprint patterns exist in 15 of 17 patients. Their EZ-Fingerprint model predicted 64 contacts and 58 of them are inside patients resected areas. By using the resection zone as ground truth, their model achieved 90.6% positive predictive value and 0.7% false-positive rate.

With the clear definition, the well shaped fingerprint can be easily identified by



(a) Wavelet image of green signal in Figure 6.1



(b) Wavelet image of red signal in Figure 6.1

Figure 6.2: Time-frequency domain representation for SEEG signals in a 40-seconds window. The x-axis is the time axis ranges from 0 - 40 seconds and the y-axis is the frequency axis ranges from 0 - 200Hz.

machine learning algorithms. The challenge is to recognize the “bad” samples as shown in Figure 6.3. The fast activity can be barely seen in Figure 6.3a and Figure 6.3b lacks pre-ictal pikes. To overcome this problem, we introduce a deep learning approach for the seizure localization based on image classification. In this study, we propose a method to locate channels inside epileptogenic zone using SEEG signals. Our major contribution includes:

- We developed a “one-click” batch processing pipeline to pre-process SEEG data, extract time-frequency features, predict results for each channel, and save output as a image.
- We built a transfer learning model using a pre-trained ResNext50 structure and trained with image augmentation.
- Our model achieved 88.22% accuracy, 34.99% sensitivity, 1.02% false positive rate, and 34.3% positive likelihood rate.

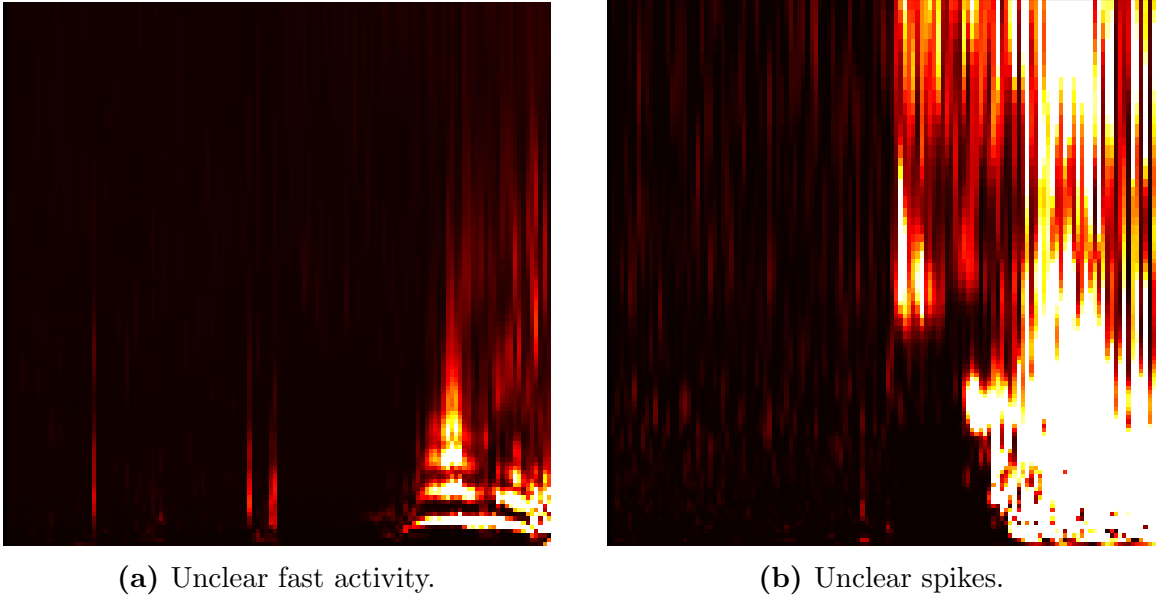


Figure 6.3: Two “bad” fingerprint examples.

6.2 Datasets

In this study, we used SEEG data of five patients from Memorial Hermann Hospital at the Texas Medical Center, which is the primary hospital affiliate of McGovern Medical School at UTHealth. Memorial Hermann Hospital at the Texas Medical Center has one of the region’s largest and most comprehensive EMU. The state-of-the-art EMU provides long-term video-EEG monitoring for adult patients. All five patients have performed craniotomies for treatment and four of five became seizure-free after their surgeries. A special case is Patient-1. The patient had two surgeries, the first resection included three channels was not fully successful because it did not avoid the seizures. After another resection with three more channels inside the area, the patient is now seizure-free. The only case with seizures remaining after resection is Patient-5.

We extracted one seizure for each subject recorded before the surgery. The start time of seizures and resection areas were marked by neurologists. The EEG signal files for processing are stored in the EDF format.

Table 6.1: The details of collected data of five patients.

Subject	sampling rate	seizure free	channels	channels in resection area
Patient-1	1000Hz	Y	172	6
Patient-2	2000Hz	Y	190	3
Patient-3	2000Hz	Y	134	47
Patient-4	2000Hz	Y	118	14
Patient-5	1000Hz	N	157	23

6.3 Pre-processing

The entire pre-processing for EEG signals is shown in Figure 6.4. The whole procedure builds our channel-based experiment dataset includes the training data and testing data. Using the input of the EDF files, we first extracted the EEG signal in 40-second segments at seizure onset. For each seizure, we also extracted a 40-second segment one minute before the extracted seizure onset segment. Then we applied artifact removal and continuous wavelet transform on every channel and normalize the data with the pre-ictal wavelet coefficients. Finally, we created the channel-based wavelet image dataset. To enlarge the training set and to enhance the robustness of the deep learning model, we generated new augmented images from each image used for training. In this section, each step is described in detail as following.

6.3.1 Channel-based Segmentation

In this study, we focus on analyzing the signals near the start of seizures and aim to build a classifier for channels with seizure activities and channels without seizure activities. The whole processing is channel-based, which means we process the EEG signal from each channel independently. Figure 6.5 shows a segmentation example on

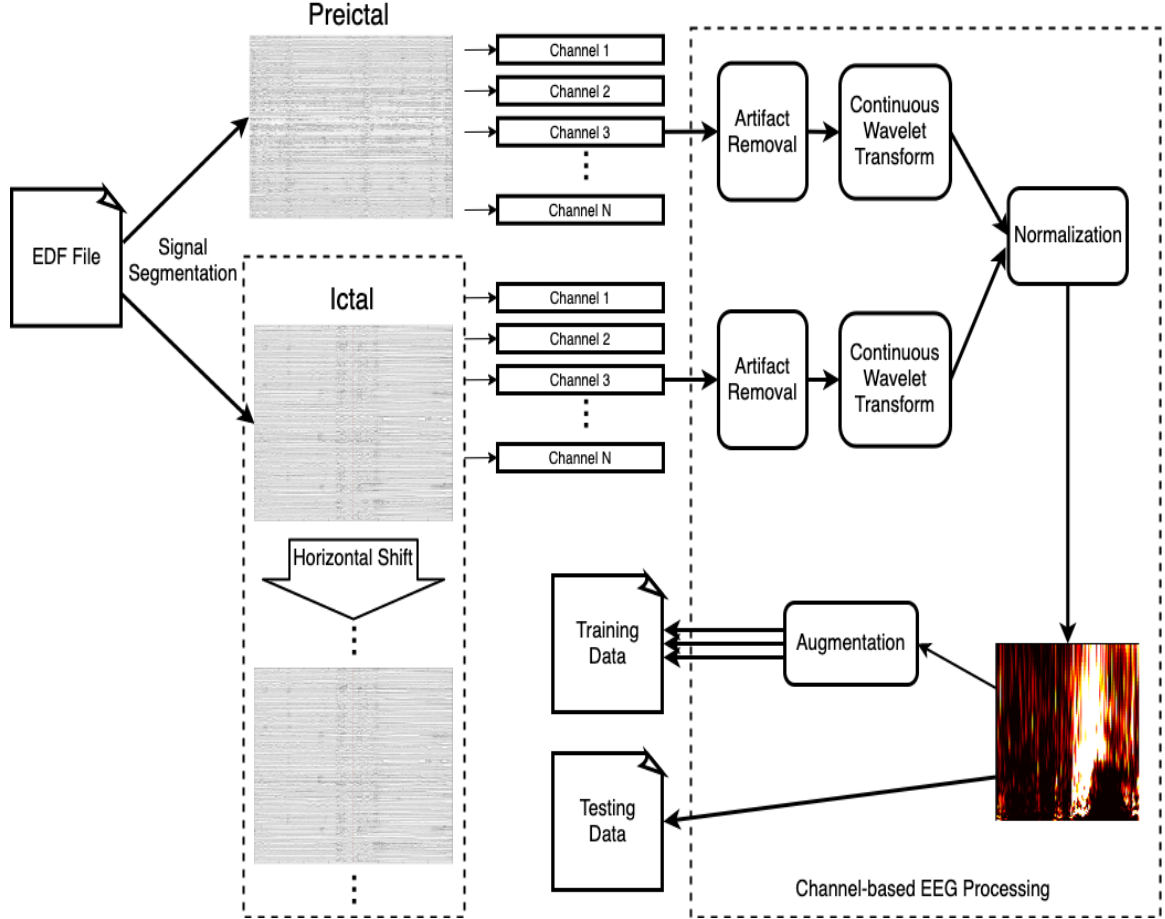


Figure 6.4: Workflow of SEEG signal data pre-processing.

channel RSMA2-RSMA3 from Patient-1. We extracted two parts from the original recording, one is the ictal signal and another is the pre-ictal. Ictal segment (in the red dotted rectangle) represents the rapid brain activity change from 20 seconds before the seizure starts and 20 seconds after the seizure starts. The pre-ictal segment (in the blue dotted rectangle) represents the background brain activity before the seizure starts, which is usually much more stable than the seizure activities. Here, we chose the pre-ictal segment from 120 seconds to 80 seconds before the seizure onset to avoid typical spikes close to the seizure. Before CWT, we implemented electrical artifact removal using a band stop filter at 60Hz, 120Hz and 180Hz.

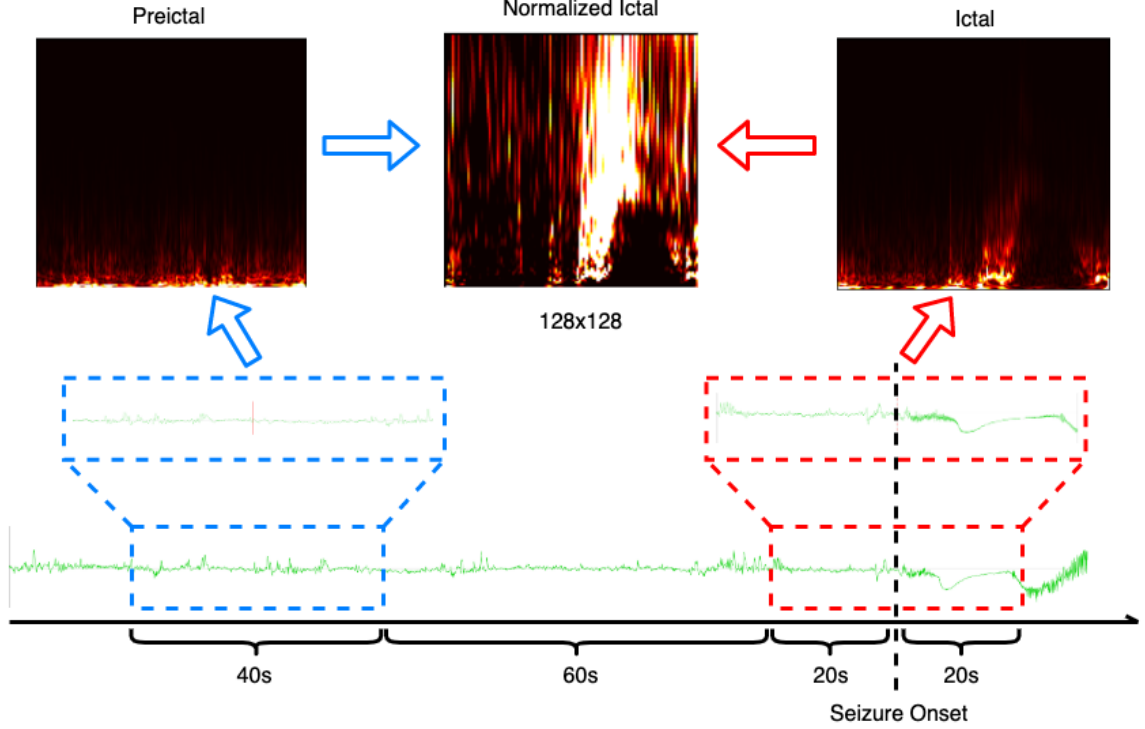


Figure 6.5: Channel-based segmentation on channel RSMA2-RSMA3 from Patient-1.

6.3.2 Time-frequency Image Generation

The next step is transforming the S EEG signal data from the time domain to the time-frequency domain, and save the transformed features as images. Unlike seizure detection and seizure prediction, seizure localization is not necessary to be a real-time task. so we used continuous wavelet transform, which is computation-consuming but provides more details than FFT and DWT. We chose Morlet wavelet and set $w0=8$ for the trade-off parameter between time resolution and frequency resolution.

A common feature for EEG data is most of the signal energy is in lower frequencies so the wavelet coefficients always have a much larger value for lower frequencies. In Figure 6.5, the left, and right wavelet images are originally transformed from signal data. The bright (high energy) areas are at the bottom of the image for both the ictal segment and the pre-ictal segment. To emphasize the change but not the absolute value of energy at different frequencies, we performed normalization of the

ictal segment with the pre-ictal segment. At frequency f , we have the normalized wavelet coefficients list for the ictal segment:

$$normalized\ ictal\ coefficient(f) = \frac{ictal\ coefficient(f) - mean(pre-ictal\ coefficient(f))}{standard\ deviation(pre-ictal\ coefficient(f))}$$

After normalization, if a significant energy change occurs at higher frequencies, it can be recognized in the time-frequency representation (the middle wavelet image in Figure 6.5). The example is from a channel inside a resected area, the processed image shows important features of seizure activities include pre-ictal spikes, quick activities, and low-frequency suppression. To better feed to the deep learning model we used in this study and to reduce the time for model training, we reshaped and down-sampled the original wavelet image with the size from 200x400 to 128x128.

6.3.3 Image Augmentation

We collected 771 channel-based samples from five seizures. 93 channels of them are inside the resected area, so we label them as positive in this study. Similarly, The data from channels that are outside the resected area are labeled as negative samples. The ratio of positive samples and negative is 1:7.29. To balance the dataset, we implemented five image augmentation methods on positive labeled data. The five methods are:

- Horizontal shift: moving all pixels of the image to left or right.
- Vertical shift: moving all pixels of the image up or down.
- Zoom in: interpolating pixel values in the center part of the image.
- Zoom out. adding new pixels values outside the image.
- Brightness reduce: darkening all pixels of the image.

Figure 6.6 shows five generated images by performing five augmentation methods on an original wavelet image. Figure 6.6 (a) is the normalized wavelet image

processed using channel RSMA2-RSMA3 of Patient-1. The horizontal shift is implemented at the segmentation step during pre-processing. As displayed in Figure 6.4, the augmented signal segments are created by sliding the ictal signal segment window to left and right. In our experiments, we used a sliding step of one second, and the total sliding length is five seconds in both directions. The other four methods are used on the processed wavelet images. All augmentation results of the four types are created by a randomly selected parameter in a range. The vertical shift range is 10% to 30% of the image height. The zoom-in method makes the center area 10% to 30% larger or closer. The zoom-out method makes the area in the original image 10% to 30% smaller or further away. The brightness reduces the method darkens the image between 20% to 50%.

Another advantage of image augmentation is the enlarged image dataset can improve the generalization ability of the model. Considering we only have five collected seizures, the total sample number is relatively small compared to a typical deep learning task. Using image augmentation, we increased the total image sample from 771 to 14,916.

6.4 Classification Model

The aim of this study is to build a classification model to automatically identify the epileptogenic zone using SEEG signals near the onset of seizures. We introduce a deep learning approach to the problem. Our model uses a wavelet RGB image with size 128x128x3 as the input, the output is a prediction to whether the signal of the image is inside of the epileptogenic zone or not. Our experiment includes two deep learning structure. The first one in Figure 6.7 is a stacked three-layer convolutional neural network. Each layer contains a convolutional layer, a max-pooling layer, and a batch normalization layer. The second structure in Figure 6.8 is using a transfer learning method as a feature extractor. The transfer learning structure is ResNext50

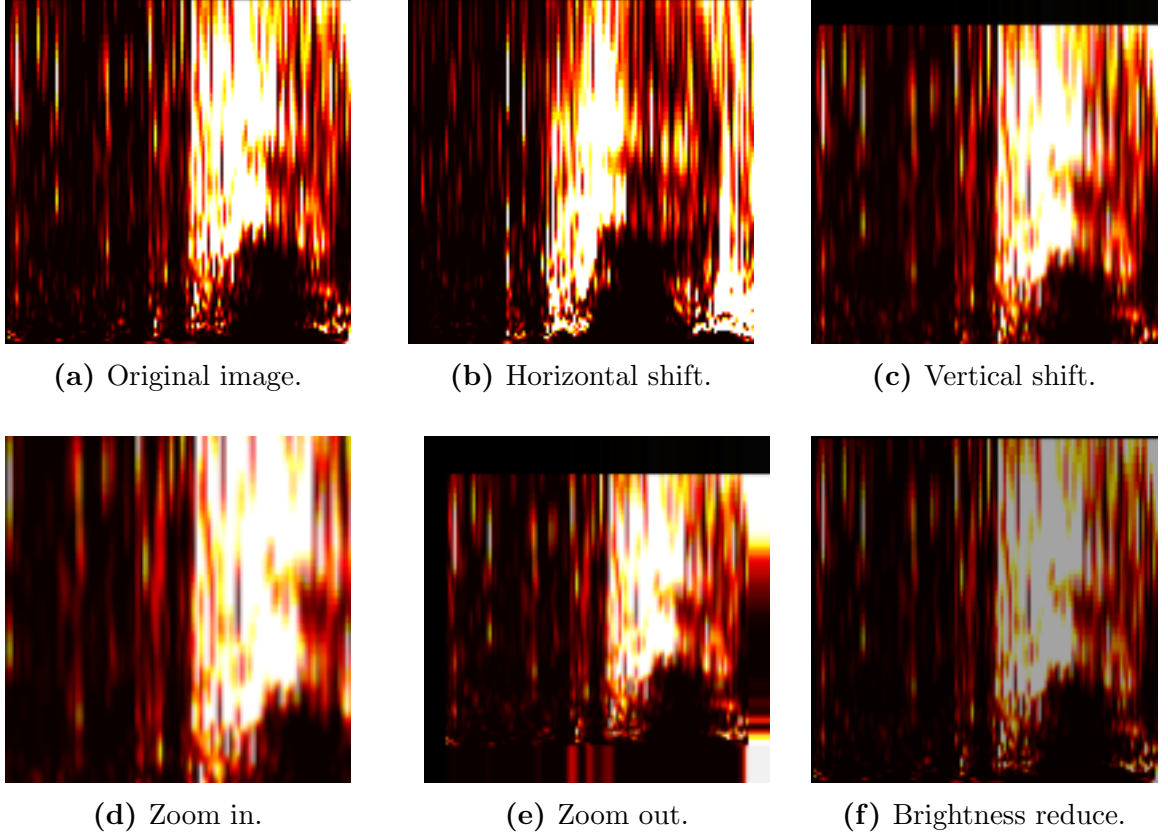


Figure 6.6: Examples of five augmentation methods performed on a original wavelet image. The image is created by signals from channel RSMA2-RSMA3 of Patient-1.

with weights pre-trained using the ImageNet dataset. Both deep learning structures are connected to two stacked fully connected layers with a two-class classification output.

For comparison, we trained a model using EZ-Fingerprint on our dataset. We ran the EZ-Fingerprint MatLab application to extract features and build an SVM model using their pipeline. We also include the pre-trained EZ-Fingerprint model introduced in [14] using seizures from 17 patients. Its testing results are shown in the next section.

We evaluated our model using the leave-one-out method to avoid data leaking

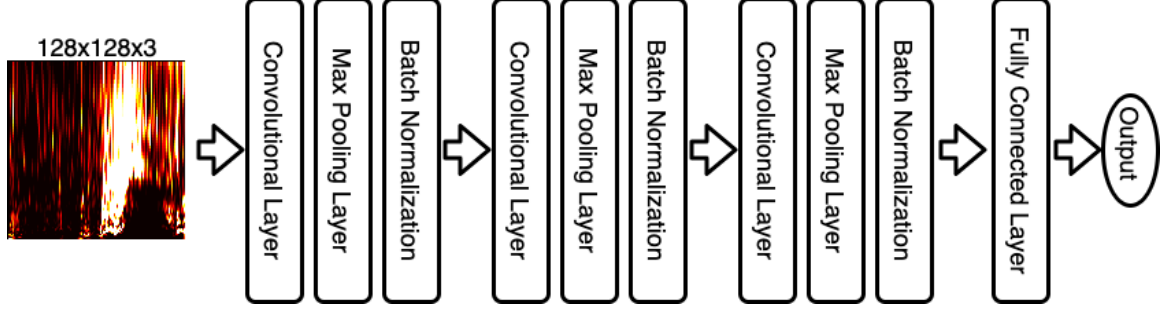


Figure 6.7: Stacked CNN model structure.

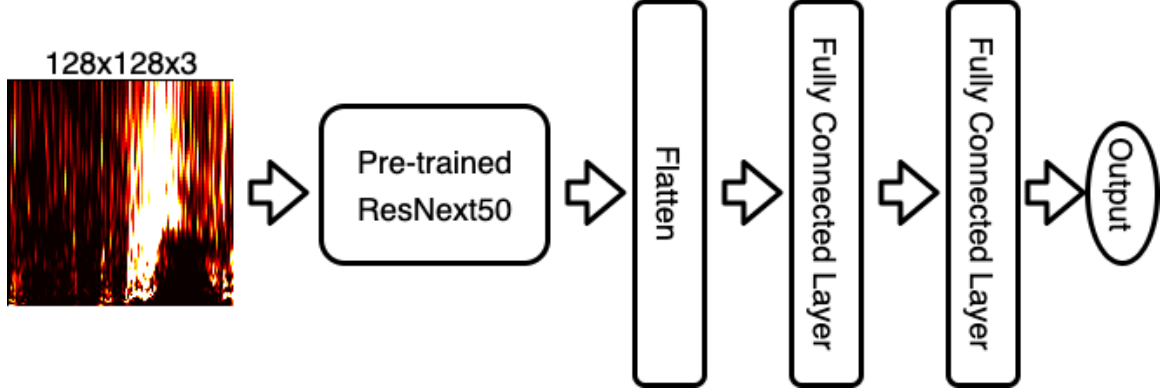


Figure 6.8: Transfer learning model structure using ResNext50.

between the training and testing set. We selected one subject for testing and the other four patients for training and repeat the process five times. The testing set for the selected subject is built using 11 segments: 1) 40 seconds segment with the seizure start time at the center of the image; 2) five left shift segments with a sliding step of 1 second; 3) five right shift segments with a sliding step of 1 second. A channel will have a prediction probability at each segment. We computed the average probability for every channel. If the probability is higher than 0.6, we count the channel as positive.

For the training set, we implemented two different labeling methods: 1) manually label positive to the image with shape similar to a fingerprint; 2) label positive to the channels inside the resected areas. We created two training sets: 1) a smaller and unbalanced dataset only includes original and horizontal shift wavelet images; 2) a larger and balanced dataset using all the augmentation methods. Besides, we add weight to positive samples in the smaller dataset. During training, we used 20% of

the training data to validate the model at the end of every epoch. The validation step can avoid the over-fitting problem by early stop when the validation loss starts increasing.

Since there is no ground truth for the epileptogenic zone, we assume that the resected area of patients is the same as the epileptogenic zone. If the model prediction is inside of the resection, we assume it is a TP (True Positive). On the contrary, if the model prediction is outside of the resection, we assume it is an FP (False Positive). We also define that the FN (False Negative) is non-epileptogenic zone prediction inside of the resection and TN (True Negative) is non-epileptogenic zone prediction outside of the resection. With the number of TP, FP, FN, and TN, we can compute four measurements for evaluation:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN}$$

$$Positive\ likelihood\ Ratio\ (PLR) = \frac{Sensitivity}{FPR}$$

We use PLR to show the combined performance of sensitivity and FPR. The higher PLR means the more increased probability for the model to find channels inside resection area. By the reason that the channel number in the resected area varies from subjects, we summary the results on multiple testing subjects by directly average the value of accuracy, sensitivity, and false-positive rate. For instance, Patient-2 has 3 channels in resection and Patient-3 has 47 channels in resection. If a model predicts 2 positives on each case, the sensitivity by definition equals $2 \times 2 / (3 + 47) = 8\%$, which does not show that the model performs well on Patient-3. Instead, we calculate the sensitivity equals to $(2/3 + 2/47)/2 = 35.46\%$.

6.5 Result

The goal of this work is to develop a deep learning model that can distinguish channels inside the epileptogenic zone from channels outside the epileptogenic zone by the time-frequency images. Table 6.2 shows the experiments results on the models using different structures, with/without augmentation and labeling methods. The model resnext50-a-fp performs best on Accuracy, FPR, and PLR. Although model cnn-na-r and cnn-a-r has better sensitivity, the two models have much higher FPR. The measurement PLR combines sensitivity and FPR to show the confidence level for positive predicted by a model.

Table 6.2: Model performance comparison. **cnn**: convolutional **neural network** model. **resnext50**: **resnext50** transfer learning model. **na**: unbalanced **non**-augmentation dataset. **a**: balanced **a**ugmentation dataset. **fp**: labeling using **f**inger**p**rint wavelet pattern. **r**: labeling using **r**esection zone.

Model	Acc (%)	Sen (%)	FPR (%)	PLR
cnn-na-fp	85.35	33.93	3.63	9.35
cnn-na-r	81.36	38.34	13.43	2.85
cnn-a-fp	86.02	31.64	4.92	6.43
cnn-a-r	80.22	36.70	16.67	2.2
resnext50-na-fp	87.65	27.19	1.73	15.72
resnext50-na-r	86.24	12.5	1.12	11.16
resnext50-a-fp	88.22	34.99	1.02	34.3
resnext50-a-r	86.74	15.35	1.73	8.87

We compared the performance between EZ-Fingerprint and this work, the results are listed in Table 6.3. The results include four patients who are seizure-free after surgeries. The EZ-Fingerprint model trained on the dataset in this study only predicted one positive for all 4 seizures, which indicates the model may underfit on the

dataset. The pre-trained EZ-Fingerprint model shows similar performance on our data to the performance in the previous study [14]. They reported 15.26% sensitivity and 0.7% FPR evaluated on their 17 patients dataset, and the same model averages 21.37% sensitivity and 1.33% FPR in this study. Our proposed model resnext50-a-fp overcomes the EZ-Fingerprint model at all aspects overall. The 34.3 PLR is also higher than the PLR of the EZ-Fingerprint model, which is $21.37/1.33 = 20.04$. Breaking down the results by subjects, our model can predict equal or more channels inside the resection zones and only performed worse on FPR for Patient-1 by predicting one more channel outside the resection area.

Table 6.3: Performance comparison between EZ-Fingerprint and this work.

Subject	Model	# Inside	# Outside	Acc (%)	Sen (%)	FPR (%)
Patient-1	EZ-FP	0	0	96.51	0	0
	EZ-FP pre-trained	3	1	97.67	50	0.6
	resnext50-a-fp	4	2	97.67	66.67	1.2
Patient-2	EZ-FP	0	0	98.42	0	0
	EZ-FP pre-trained	1	7	95.26	33.33	3.74
	resnext50-a-fp	1	0	98.95	33.33	0
Patient-3	EZ-FP	0	0	64.93	0	0
	EZ-FP pre-trained	1	0	65.67	2.13	0
	resnext50-a-fp	2	0	66.42	4.26	0
Patient-4	EZ-FP	1	0	88.98	7.14	0
	EZ-FP pre-trained	0	1	87.29	0	0.96
	resnext50-a-fp	5	3	89.83	35.71	2.88
Average	EZ-FP	0.25	0	87.21	1.785	0
	EZ-FP pre-trained	1.25	2.25	86.47	21.37	1.33
	resnext50-a-fp	3	1.25	88.22	34.99	1.02

On the four seizure-free patients, our proposed model gave 17 positive (sum of

numbers inside left circle in 6.9) and EZ-Fingerprint gave 14 positive (sum of numbers inside right circle in 6.9). The two methods have an agreement on five positive predictions. In the five channels, four (80%) of them are inside the resection zone. Based on the current dataset, we can say if a channel is predicted as positive by both of the machine learning methods, it has a high probability to be inside the epileptogenic zone. Only one of the disagreed positive predictions made by EZ-Fingerprint is inside the resection zone while our model predicted eight more channels inside the resection area than EZ-Fingerprint did. The difference shows the deep learning model can catch more features from the time-frequency images so it can recognize more true positive samples. As shown in the bottom half of 6.9, the EZ-Fingerprint model has two times false positive on disagreed predictions. Because all of the four patients are seizure-free, the false-positive channels should not be inside epileptogenic zones.

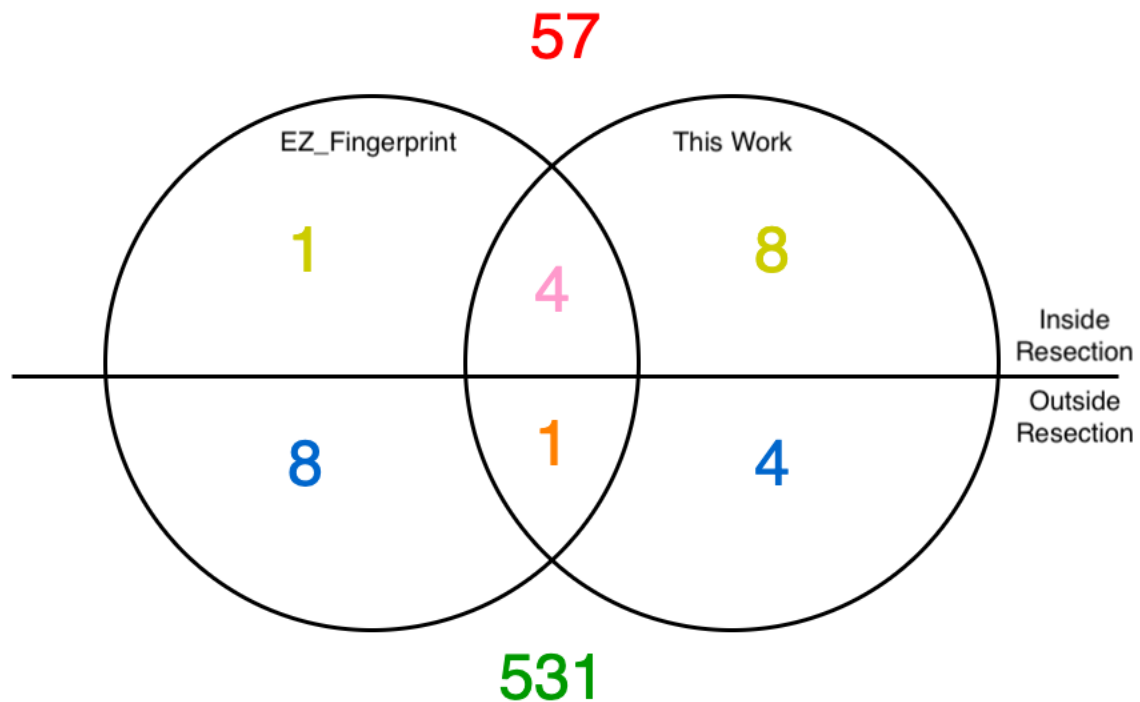


Figure 6.9: Venn diagram of overall performance for EZ-Fingerprint and this work.

6.6 Discussion

Clinical significance. In this study, we introduced an epileptogenic zone localization methods using SEEG signals leveraging wavelet transform and deep learning. Existing clinical method requires multiple tests and presurgical evaluations to estimate the epileptogenic zone for resection. Our method together with multiple approaches may contribute to decide and reduce the resection area. Our model shows a 34.99% recall on electrodes in the resected area. The resected area is larger than or equal to the epileptogenic zone for a seizure-free patient, so the real recall may be higher than the results of the experiment. In our dataset, Patient-1 had two surgeries, the first one with three resected channels was not successful, and the second one with another three channels leads the patient to seizure-free. We can claim that the second resection contains a more important area that causes seizures. According to the results, we found the prediction agreement by both models has a high probability inside the epileptogenic zone. In the Patient-1 case, two of four agreed channels are inside the resection zone of the second surgery, one in the resection zone of the first surgery and one outside both resection areas. It indicates that our model made a better choice of resection than the real resected area in the first surgery. If the results of our model were considered, the success rate of the surgeries may be increased. For case Patient-5 in 6.4, the two outside resection agreements may include the real epileptogenic zone but need further proof.

Usability and running time. A Matlab-based application using the algorithm in [14] had been released to the public. The whole pipeline can be processed using a graphic user interface so it is friendly to neurologists. However, the program includes many manual works like selecting the start and endpoint of quick activities. Our pipeline is more straight forward because the deep learning model extracts the from

Table 6.4: Case study on Patient-1 and Patient-5.

Subject	Model	# inside	# outside	# resection	# outside agreement
Patient-1	EZ-FP pre-trained	1	3	3	3
Surgery 1	resnext50-a-fp	1	5		
Patient-1	EZ-FP pre-trained	3	1	6	1
Surgery 1&2	resnext50-a-fp	4	2		
Patient-5	EZ-FP pre-trained	0	24	23	2
	resnext50-a-fp	0	3		

the images automatically. Besides, our program is faster on the pre-processing. The running speed performance comparison is shown in 6.5. Our pipelines differ on the order of processing: our pipeline implements artifact removal before CWT but EZ-Fingerprint does in a reversed way.

Table 6.5: Process running time of the two pipelines. The testing machine is Mac-Book Pro 2015 with 2.7GHz Intel Core i5 CPU, and 8GB RAM.

Process	EZ-Fingerprint	This work
CWT	709 seconds	437 seconds
CWT + ICA Artifact Removal	5488 seconds	511 seconds

Dataset size of this work. A limitation of this work is the size of the dataset. We reported our preliminary experiments on five subjects. Future work will include more SEEG data. Because seizure signal varies from subjects, more subjects involved in the study can improve the performance and generalization ability of our deep learning model.

6.7 Conclusion

In this study, we developed a “one-click” batch processing pipeline to pre-process SEEG data, extract time-frequency features, predict the results for each channel, and

save the output as an image. The core of this work is a deep learning method for seizure localization based on image classification. Our experiment shows a transfer learning model performs best. The model uses a pre-trained ResNext50 structure and was trained with an image augmentation dataset labeling by fingerprint. Our model achieved 88.22% accuracy, 34.99% sensitivity, 1.02% false-positive rate, and 34.3% positive likelihood rate. The results indicate an improvement in seizure localization from previous work.

CHAPTER 7. Conclusions and Future Work

In the last chapter, we first summarize our work in the dissertation and then conclude our contribution to each chapter. At last, we provide possible improvements in our methods and potential direction for future work.

7.1 Conclusions

Our goal of this dissertation is to provide a time-saver solution for current epilepsy researchers. We also report our experiments on EEG data analysis on a large cross-site database. We developed an end-to-end pipeline for the machine learning approach on EEG signal classification. The multi-site input data is the CSR epilepsy data collected from seven individual epilepsy centers. In Chapter 3, we described how our system extracts epilepsy temporal information from EMU patient reports, EEG signal files in EDF format, and annotation text files. The system includes a graphical interface for temporal data query and visualization for epilepsy cohort discovery. In Chapter 4, we introduced an automatic channel-based cross-patient seizure detection model and evaluated it on two scalp EEG datasets using continuous long-term EEG signals. In Chapter 5, we built a transfer learning model for patient-specific real-time seizure prediction. Finally, in Chapter 6, we provided a deep learning approach for the cross-patient seizure localization model leveraging the epileptogenic zone fingerprint technique. Comparing to existing methods, our work contributes to the following aspects:

Large-scale temporal information extraction. Our ontology-guided multi-site epilepsy temporal information system processed 2,497 epilepsy patients with 3,169 reports from 7 epilepsy centers across the U.S. and Europe. We extracted 451,076 temporal annotations from 42,239 EEG files. Moreover, the prospective system is scalable for new incoming epilepsy data and new independent sites. We constructed

vocabulary sets including 46 standard annotation terms for ontological annotation elements and fuzzy matched 6,687 annotations for high-quality queries. Annotation data, EDF header data, and patient report data make the temporal data system a comprehensive information database for epilepsy study.

Prospective temporal data quality measurements. Our system prospectively integrates the epilepsy temporal data in CSR once a week. We automatically calculated the data quality measurements for the epilepsy temporal data. The results from September 4th, 2019 version show the CSR dataset has 99.12% annotation completeness, 61.71% EEG signals completeness for all existing monitoring, and 0.85% signal file duplication rate. The phenotypic data quality measurement shows improvement from an older version of February 2016, which indicates the measurements can enhance the data quality of a prospective dataset.

Graphical temporal query. Our system provides a web-based temporal query interface developed by the RoR development framework, which provides the first graphical temporal query system in the epilepsy research area. Both of our query widget and results representation are displayed in the graphical timeline. Users can build a query by creating time points and intervals of annotation and drag them to demand order. The temporal query canvas can generate all 13 Allen’s interval algebra with minimal user intervention. By using our interface, users can download the query results in the CSV format for preliminary research or building their datasets, which is a fast and accurate solution for cohort discovery and pattern discovery on large-scaled CSR epilepsy data.

EEG signal classification performance. We developed three machine learning models for EEG signal classification. For seizure detection, we extracted 135 features from each 8-second channel signals using signal processing methods like DWT, autocorrelation, PSD, and FFT and used such features with labels to train an XG-Boost model. The model is evaluated on 195 seizure cases and 98.73% of the non-

seizure period, which is significantly larger than the testing of existing work. The case-based testing results show that our model detected 90.75% lead seizures, and achieved 92.23% overall accuracy, 93.57% specificity, 81.08% sensitivity, and 85.41% AUC in the segment-based evaluation. Our seizure prediction approach consists of a multi-channel transfer learning model and a prediction decision queue using multi-channel voting. A seizure prediction evaluation was performed using 53 pre-ictal EEG signal segments and 265 hours inter-ictal EEG data. Our results show that after 25 epochs training, the transfer learning model using ResNeXt50 pre-trained model reached 86.79% sensitivity and 3.38% false-positive rate. We also developed a “one-click” batch processing pipeline for seizure localization. The model uses a pre-trained ResNext50 structure and was trained with an image augmentation dataset labeling by fingerprint. Our model achieved 88.22% accuracy, 34.99% sensitivity, 1.02% false-positive rate, and 34.3% positive likelihood rate. The results indicate an improvement in seizure localization from previous work.

Continuous long-term EEG evaluation. Our work provides a new evaluation method using continuous long-term EEG recordings leveraging CSR large-scale data volume. By using the data quality measurement, we built a sub-dataset that has high data coverage with a 98.98% EEG signal completeness which is significantly greater than the 59.57% data completeness of CHB-MIT scalp EEG dataset. Our evaluation results were based on 2,097 hours testing for our seizure detection model and 1,506 hours for our seizure prediction model. Our more completed dataset is more ideal to simulate a real-world situation so the evaluation of our dataset is more convincing. The evaluation results on the continuous long-term EEG data can truly reflect the model performance during daily use, which can be a benchmark for future seizure detection methods on EEG signals.

7.2 Future Work

In this section, we describe the limitation of our work in the dissertation and possible solutions for improvement.

7.2.1 A More Powerful Cloud-based EEG interface

In Chapter 3, we developed an interface that can provide a graphics-based annotation visualization for epilepsy temporal data. The web application allows users to access the large-scaled SCR data without installing any other software and downloading any data. Our work has potential to be combined with other cloud-based EEG visualization tools [5] and MRI readers. The more powerful interface will be convenient for EEG signal exploration and for experts to curate the annotations. Another aspect is that a software can be developed for automatically auditing the existing errors to improve data quality. From the measurements, we can locate the potential errors, but manual curation is still time-consuming. A potential approach can be creating new ontological annotations or curating existing free-text annotations to standard terminology using natural language processing methods. Moreover, our seizure detection algorithm has shown good performance on detecting the seizure activities, so the model has the potential to automatically mark seizure onset and end annotation, but further experiments are needed.

7.2.2 EEG Signal Quality Assurance

In Chapter 3, we introduced data quality measurements for epilepsy temporal data. For EEG signals, we only measured the length of each EDF file and the total coverage of each long-term monitoring. Future work of the temporal data quality can be done regarding the EEG signal quality. The EEG signal, especially scalp EEG, may contain certain artifacts that can be hardly removed like muscle activity and eye movement.

In a worse situation, part of EEG signals are damaged so such a period of data does not contain any valuable information. Measuring the signal quality is crucial for future EEG analysis to avoid the “Garbage in garbage out” problem. It is possible to build an EEG signal classifier to recognize bad signals. Existing methods [68] provide multiple solutions for different purposes but lack of evaluation on long-term real EEG signal. Because labeling artifacts manually is labor-consuming, an automatic artifacts identifier or removal will be essential for the improvement of EEG signal quality.

7.2.3 Features Enhancement on EEG Signal Analysis

The machine learning methods we described in Chapter 4 and 6 are trained with the cross-patient dataset. Since brain activities vary from subjects, considering the subjects’ variance is a potential point to improve the performance of classifiers. With the increment of hardware’s computational ability, more data can be used in future research. We can enhance the feature by using more data other than signals, for instance, patients’ phenotypical data, seizure types, seizure frequency, etc. But with more features used, the dataset will be divided into smaller groups with fewer samples. Future work will investigate the performance of new machine learning approaches like few-shot learning and meta-learning. Another new approach for EEG signal feature extraction uses a discrete Pade spectrogram to find new brain wave patterns [69]. For projects such as seizure localization, the number of data samples may be the bottleneck for deep learning models. Using the spatial features as the location of electrodes to generate a larger set of combinations of channels can increase the dataset size. When the dimensions of feature increases, the feature selection methods can be also implemented to filter out equivalent features and bad features.

REFERENCES

- [1] Matthew M Zack and Rosemarie Kobau. National and state estimates of the numbers of adults and children with active epilepsy—united states, 2015. *MMWR. Morbidity and mortality weekly report*, 66(31):821, 2017.
- [2] Guo-Qiang Zhang, Licong Cui, Samden Lhatoo, Stephan U Schuele, and Satya S Sahoo. Medcis: multi-modality epilepsy data capture and integration system. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1248. American Medical Informatics Association, 2014.
- [3] Ernst Niedermeyer and FH Lopes da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.
- [4] Richard W Homan, John Herman, and Phillip Purdy. Cerebral location of international 10–20 system electrode placement. *Electroencephalography and clinical neurophysiology*, 66(4):376–382, 1987.
- [5] Xiaojin Li, Yan Huang, Shiqiang Tao, Licong Cui, Samden D Lhatoo, and G Zhang. Seizurebank: A repository of analysis-ready seizure signal data. In *AMIA Annual Symposium Proceedings*, volume 2019, pages 1111–1120, 2019.
- [6] Jean Gotman. Automatic recognition of epileptic seizures in the eeg. *Electroencephalography and clinical Neurophysiology*, 54(5):530–540, 1982.
- [7] Yang Li, Xu-Dong Wang, Mei-Lin Luo, Ke Li, Xiao-Feng Yang, and Qi Guo. Epileptic seizure classification of eegs using time–frequency analysis based multiscale radial basis functions. *IEEE journal of biomedical and health informatics*, 22(2):386–397, 2017.
- [8] Hojjat Adeli, Ziqin Zhou, and Nahid Dadmehr. Analysis of eeg records in an epileptic patient using wavelet transform. *Journal of neuroscience methods*, 123(1):69–87, 2003.
- [9] Kemal Polat and Salih Güneş. Classification of epileptiform eeg using a hybrid system based on decision tree classifier and fast fourier transform. *Applied Mathematics and Computation*, 187(2):1017–1026, 2007.
- [10] Yun Park, Lan Luo, Keshab K Parhi, and Theoden Netoff. Seizure prediction with spectral power of eeg using cost-sensitive support vector machines. *Epilepsia*, 52(10):1761–1770, 2011.
- [11] Md Mursalin, Yuan Zhang, Yuehui Chen, and Nitesh V Chawla. Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. *Neurocomputing*, 241:204–214, 2017.

- [12] Emina Alickovic, Jasmin Kevric, and Abdulhamit Subasi. Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction. *Biomedical signal processing and control*, 39:94–102, 2018.
- [13] Steven N Baldassano, Benjamin H Brinkmann, Hoameng Ung, Tyler Blevins, Erin C Conrad, Kent Leyde, Mark J Cook, Ankit N Khambhati, Joost B Wagenaar, Gregory A Worrell, et al. Crowdsourcing seizure detection: algorithm development and validation on human implanted device recordings. *Brain*, 140(6):1680–1691, 2017.
- [14] Olesya Grinenko, Jian Li, John C Mosher, Irene Z Wang, Juan C Bulacio, Jorge Gonzalez-Martinez, Dileep Nair, Imad Najm, Richard M Leahy, and Patrick Chauvel. A fingerprint of the epileptogenic zone in human epilepsies. *Brain*, 141(1):117–131, 2018.
- [15] Mark J Cook, Terence J O’Brien, Samuel F Berkovic, Michael Murphy, Andrew Morokoff, Gavin Fabinyi, Wendyl D’Souza, Raju Yerra, John Archer, Lucas Litewka, et al. Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. *The Lancet Neurology*, 12(6):563–571, 2013.
- [16] Isabell Kiral-Kornek, Subhrajit Roy, Ewan Nurse, Benjamin Mashford, Philippa Karoly, Thomas Carroll, Daniel Payne, Susmita Saha, Steven Baldassano, Terence O’Brien, et al. Epileptic seizure prediction using big data and deep learning: toward a mobile system. *EBioMedicine*, 27:103–111, 2018.
- [17] Lindsay F Haas. Hans berger (1873–1941), richard caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(1):9–9, 2003.
- [18] Eeg / erp data available for free public download. https://sccn.ucsd.edu/~arno/fam2data/publicly_available_EEG_data.html. Accessed: Dec 23, 2018.
- [19] A list of all public eeg-datasets. <https://github.com/meagmohit/EEG-Datasets>. Accessed: Dec 23, 2018.
- [20] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.
- [21] Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

- [22] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.
- [23] Chb-mit scalp eeg database. <https://www.physionet.org/pn6/chbmit/>. Accessed: Dec 23, 2018.
- [24] American epilepsy society seizure prediction challenge. <https://www.kaggle.com/c/seizure-prediction>. Accessed: Dec 23, 2018.
- [25] Rohit Shankar, Elizabeth J Donner, Brendan McLean, Lina Nashef, and Torbjörn Tomson. Sudden unexpected death in epilepsy (sudep): what every neurologist should know. *Epileptic Disorders*, 19(1):1–9, 2017.
- [26] The center for sudep research. <http://sudepresearch.org/>. Accessed: April 20, 2020.
- [27] Samden D Lhatoo, Maromi Nei, Manoj Raghavan, Michael Sperling, Bilal Zonjy, Nuria Lacuey, and Orrin Devinsky. Nonseizure sudep: sudden unexpected death in epilepsy without preceding epileptic seizures. *Epilepsia*, 57(7):1161–1168, 2016.
- [28] Bob Kemp, Alpo Värri, Agostinho C Rosa, Kim D Nielsen, and John Gade. A simple format for exchange of digitized polygraphic recordings. *Electroencephalography and clinical neurophysiology*, 82(5):391–393, 1992.
- [29] Burcu Yildiz and Silvia Miksch. Ontology-driven information systems: Challenges and requirements. In *International Conference on Semantic Web and Digital Libraries. Indian Statistical Institute Platinum Jubilee Conference Series*, pages 35–44. Citeseer, 2007.
- [30] Satya S Sahoo, Samden D Lhatoo, Deepak K Gupta, Licong Cui, Meng Zhao, Catherine Jayapandian, Alireza Bozorgi, and Guo-Qiang Zhang. Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care. *Journal of the American Medical Informatics Association*, 21(1):82–89, 2014.
- [31] Nickoal Eichmann-Kalwara. The state of open data report 2018. *The Idealis*, 2018.
- [32] Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/>. Accessed: March, 23, 2018.

- [33] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.
- [34] Jessica S Ancker, Sarah Shih, Mytri P Singh, Andrew Snyder, Alison Edwards, Rainu Kaushal, Hitec Investigators, et al. Root causes underlying challenges to secondary use of data. In *AMIA Annual Symposium Proceedings*, volume 2011, page 57. American Medical Informatics Association, 2011.
- [35] Licong Cui, Yan Huang, Shiqiang Tao, Samden D Lhatoo, and Guo-Qiang Zhang. Odacci: Ontology-guided data curation for multisite clinical research data integration in the ninds center for sudep research. In *AMIA Annual Symposium Proceedings*, volume 2016, page 441. American Medical Informatics Association, 2016.
- [36] Wolfgang Dorda, Walter Gall, and Georg Duftschmid. Clinical data retrieval: 25 years of temporal query management at the university of vienna medical school. *Methods of information in medicine*, 41(02):89–97, 2002.
- [37] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [41] Wenting Tu and Shiliang Sun. A subject transfer framework for eeg classification. *Neurocomputing*, 82:109–116, 2012.
- [42] Adam Page, Colin Shea, and Tinoosh Mohsenin. Wearable seizure detection using convolutional neural networks with transfer learning. In *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1086–1089. IEEE, 2016.
- [43] Yizhang Jiang, Dongrui Wu, Zhaohong Deng, Pengjiang Qian, Jun Wang, Guanjin Wang, Fu-Lai Chung, Kup-Sze Choi, and Shitong Wang. Seizure classification from eeg signals using transfer learning, semi-supervised learning and tsf fuzzy system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(12):2270–2284, 2017.

- [44] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014.
- [45] Tomáš Mikolov, Anoop Deoras, Daniel Povey, Lukáš Burget, and Jan Černocký. Strategies for training large scale neural network language models. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 196–201. IEEE, 2011.
- [46] Tomáš Mikolov, Anoop Deoras, Daniel Povey, Lukáš Burget, and Jan Černocký. Strategies for training large scale neural network language models. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 196–201. IEEE, 2011.
- [47] Paul Fergus, A Hussain, David Hignett, Dhiya Al-Jumeily, Khaled Abdel-Aziz, and Hani Hamdan. A machine learning system for automated whole-brain seizure detection. *Applied Computing and Informatics*, 12(1):70–89, 2016.
- [48] Pierre Thodoroff, Joelle Pineau, and Andrew Lim. Learning robust features using deep learning for automatic seizure detection. In *Machine learning for healthcare conference*, pages 178–190, 2016.
- [49] Chulkyun Park, Gwangho Choi, Junkyung Kim, Sangdeok Kim, Tae-Joon Kim, Kyeongyuk Min, Ki-Young Jung, and Jongwha Chong. Epileptic seizure detection for multi-channel eeg with deep convolutional neural network. In *2018 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–5. IEEE, 2018.
- [50] Serkan Kiranyaz, Turker Ince, Morteza Zabihi, and Dilek Ince. Automated patient-specific classification of long-term electroencephalography. *Journal of biomedical informatics*, 49:16–31, 2014.
- [51] Guangxu Xun, Xiaowei Jia, and Aidong Zhang. Detecting epileptic seizures with electroencephalogram via a context-learning model. *BMC medical informatics and decision making*, 16(2):70, 2016.
- [52] Ye Yuan, Guangxu Xun, Fenglong Ma, Qiuling Suo, Hongfei Xue, Kebin Jia, and Aidong Zhang. A novel channel-aware attention framework for multi-channel eeg seizure detection via multi-view deep learning. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 206–209. IEEE, 2018.
- [53] Xiaobin Tian, Zhaohong Deng, Wenhao Ying, Kup-Sze Choi, Dongrui Wu, Bin Qin, Jun Wang, Hongbin Shen, and Shitong Wang. Deep multi-view feature learning for eeg-based epileptic seizure detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(10):1962–1972, 2019.

- [54] Mengni Zhou, Cheng Tian, Rui Cao, Bin Wang, Yan Niu, Ting Hu, Hao Guo, and Jie Xiang. Epileptic seizure detection based on eeg signals and cnn. *Frontiers in neuroinformatics*, 12:95, 2018.
- [55] Kais Gadhoudi, Jean-Marc Lina, Florian Mormann, and Jean Gotman. Seizure prediction for therapeutic devices: A review. *Journal of neuroscience methods*, 260:270–282, 2016.
- [56] Hans O Lüders, Imad Najm, Dileep Nair, Peter Widdess-Walsh, and William Bingman. The epileptogenic zone: general principles. *Epileptic disorders*, 8(2):1–9, 2006.
- [57] Lara Jehi. The epileptogenic zone: concept and definition. *Epilepsy currents*, 18(1):12–16, 2018.
- [58] Maciej A Mazurowski, Piotr A Habas, Jacek M Zurada, Joseph Y Lo, Jay A Baker, and Georgia D Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, 21(2-3):427–436, 2008.
- [59] Ruby on rails. <https://guides.rubyonrails.org/>. Accessed: Dec 23, 2018.
- [60] A Hunt and D Thomas. The pragmatic programmer: From journeyman to master, 1999.
- [61] Tensorflow. <https://www.tensorflow.org/>. Accessed: Dec 23, 2018.
- [62] Georgios Gousios, Bogdan Vasilescu, Alexander Serebrenik, and Andy Zaidman. Lean ghtorrent: Github data on demand. In *Proceedings of the 11th working conference on mining software repositories*, pages 384–387, 2014.
- [63] Epilepsy - world health organization. <https://www.who.int/news-room/fact-sheets/detail/epilepsy>. Accessed: April 10, 2020.
- [64] Mary Jane England, Catharyn T Liverman, Andrea M Schultz, and Larisa M Strawbridge. A summary of the institute of medicine report: epilepsy across the spectrum: promoting health and understanding. *Epilepsy & behavior: E&B*, 25(2):266, 2012.
- [65] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [66] Nhan Duy Truong, Anh Duy Nguyen, Levin Kuhlmann, Mohammad Reza Bonyadi, Jiawei Yang, and Omid Kavehei. A generalised seizure prediction with convolutional neural networks for intracranial and scalp electroencephalogram data analysis. *arXiv preprint arXiv:1707.01976*, 2017.

- [67] Texas comprehensive epilepsy program. <https://med.uth.edu/neurosurgery/mischer-neuroscience-institute/epilepsy-program/>. Accessed: April 20, 2020.
- [68] Malik Muhammad Naeem Mannan, Muhammad Ahmad Kamran, and Myung Yung Jeong. Identification and removal of physiological artifacts from electroencephalogram signals: A review. *IEEE Access*, 6:30630–30652, 2018.
- [69] Luca Perotti, Justin DeVito, Daniel Bessis, and Yuri Dabaghian. Discrete structure of the brain rhythms. *Scientific reports*, 9(1):1–7, 2019.

Vita

Personal Information

- Name: Yan Huang

Education

- M.S. in Computer Science, Case Western Reserve University
Cleveland, Ohio, Jan 2019
- B.S. in Computer Science, University of Missouri
Columbia, Missouri, May 2014
- B.E. in Electrical Engineer, Shanghai University
Shanghai, China, Jul 2011

Professional Experience

- Visiting Student Trainee, University of Texas Health Science Center at Houston
Houston, Texas. Aug 2019 - May 2020
- Research Assistant, University of Kentucky
Lexington, Kentucky. Jan 2016 - May 2020
- Research Assistant, Case Western Reserve University
Cleveland, Ohio. Jun 2015 - Jan 2016
- Electrical Engineer, Oriental Cable Network Co Ltd
Shanghai, China. Jul 2011 - Jan 2012

Publications

1. Li X, Tao S, Jamal-Omidi S, Huang Y, Lhatoo SD, Zhang GQ, Cui L. Detection of Postictal Generalized Electroencephalogram Suppression: Random Forest Approach. *JMIR Medical Informatics*. 2020;8(2):e17061.
2. Li X, Huang Y, Tao S, Cui L, Lhatoo SD, Zhang GQ. SeizureBank: A Repository of Analysis-ready Seizure Signal Data. In *AMIA Annual Symposium Proceedings 2019*.
3. Huang Y, Li X, Tao S, Guo-Qiang Z, A Transfer Learning Approach to Real-time Seizure Prediction. In *Rice Data Science Conference Proceedings 2019*
4. Yao X, Li X, Ye Q, Huang Y, Cheng Q, Zhang GQ. A robust deep learning approach for automatic classification of seizures against non-seizures. arXiv preprint arXiv:1812.06562. 2018 Dec 17.

5. Guo-Qiang Z, Yan H, Licong C. Can SNOMED CT Changes Be Used as a Surrogate Standard for Evaluating the Performance of Its Auditing Methods?. In *AMIA Annual Symposium Proceedings 2017* (Vol. 2017, p. 1903).
6. Cui L, Huang Y, Tao S, Lhatoo SD, Zhang GQ. ODaCCI: Ontology-guided Data Curation for Multisite Clinical Research Data Integration in the NINDS Center for SUDEP Research. In *AMIA Annual Symposium Proceedings 2016* (Vol. 2016, p. 441).