การตรวจจับอัตโนมัติของจุดเริ่มและจุดสิ้นสุดของการชักในสัญญาณคลื่นไฟฟ้าสมอง

นายภูมิพัฒน์ บุญยกิตานนท์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

AUTOMATIC DETECTION OF EPILEPTIC SEIZURE ONSET AND OFFSET IN EEG

Mr. Poomipat Boonyakitanont

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Electrical Engineering
Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2020

Thesis Title	AUTOMATIC DETECTION OF EPILEPTIC SEIZURE ONSET AND OFFSET IN EEG
By	Mr. Poomipat Boonyakitanont
Field of Study	Electrical Engineering
Thesis Advisor	Associate Professor Jitkomut Songsiri, Ph.D.
Thesis Co advisor	Assistant Professor Apiwat Lek-uthai, Ph.D.
Thesis Co advisor	Assistant Professor Krisnachai Chomtho, M.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctor of Philosophy

Dean of the Faculty of Engineering

. . . . . .
(Professor Supot Teachavorasinskun, D.Eng.)

DISSERTATION COMMITTEE

. . . . . Chairman
(Associate Professor Supatana Auethavekiat, Ph.D.)

. . . . . Thesis Advisor
(Associate Professor Jitkomut Songsiri, Ph.D.)

. . . . . Thesis Co-advisor
(Assistant Professor Apiwat Lek-uthai, Ph.D.)

. . . . . Thesis Co-advisor
(Assistant Professor Krisnachai Chomtho, M.D.)

. . . . . Examiner
(Associate Professor Chedsada Chinrungrueng, Ph.D.)

. . . . . Examiner
(Associate Professor Setha Pan-ngum, Ph.D.)

. . . . . External Examiner
(Pasin Isarasena na Ayudhya, Ph.D.)

ภูมิพัฒน์ บุญยกิตานนท์: การตรวจจับอัตโนมัติของจุดเริ่มและจุดสิ้นสุดของการชักในสัญญาณคลื่นไฟฟ้าสมอง. (AUTOMATIC DETECTION OF EPILEPTIC SEIZURE ONSET AND OFFSET IN EEG)
อ.ที่ปรึกษาวิทยานิพนธ์หลัก : รศ. ดร.จิตโกมุท ส่งศิริ,
อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : ผศ. ดร.อภิวัฒน์ เล็กอุทัย,
อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : ผศ. นพ.กฤษณชัย ชมโท, 89 หน้า.

วิทยานิพนธ์นี้นำเสนอวิธีการตรวจจับอัตโนมัติของจุดเริ่มและจุดสิ้นสุดของการชักในสัญญาณคลื่นไฟฟ้าสมองบนหนังศีรษะเพื่อช่วยลดภาระงานจากผู้เชี่ยวชาญ ในงานนี้ การตรวจจับจุดเริ่มและจุดสิ้นสุดของการชักประกอบไปด้วยการตรวจจับการชัก และการระบุจุดเริ่มและจุดสิ้นสุดของการชัก โครงข่ายประสาทเทียมแบบสังวัฒนาการที่ปราศจากการสกัดคุณลักษณะได้ถูกออกแบบมาเพื่อตรวจจับการชักในสัญญาณคลื่นไฟฟ้าสมอง เราได้นำเสนอวิธีการประมวลผลภายหลังสองวิธี ชื่อ วิธีอิงการนับและสกอร์เน็ต เพื่อระบุจุดเริ่มและจุดสิ้นสุดจากการชักที่ตรวจจับได้ วิธีอิงการนับถูกออกแบบมาจากคุณลักษณะทางการแพทย์ของการชักที่มีการเปลี่ยนแปลงของสัญญาณคลื่นไฟฟ้าสมองในระหว่างการชักอย่างต่อเนื่อง และสกอร์เน็ตเป็นแบบจำลองอิงโครงข่ายประสาทเทียมที่ประยุกต์วิธีอิงการนับ โดยการเพิ่มความยืดหยุ่นเข้าไปเพื่อให้แบบจำลองมีความจำเพาะน้อยลง เราได้เสนอฟังก์ชันสูญเสียแบบไดซ์เชิงลอการิทึม เพื่อเป็นฟังก์ชันวัตถุประสงค์สำหรับการเรียนรู้ของสกอร์เน็ตในการจัดการปัญหาความไม่สมมดุลของข้อมูล ในวิทยานิพนธ์นี้ เราได้ทดสอบวิธีที่นำเสนอกับฐานข้อมูลสาธารณะด้วยแผนแบบเฉพาะคนไข้ผลปรากฏว่า โดยภาพรวมแล้ว โครงข่ายประสาทเทียมแบบสังวัฒนาการสามารถตรวจจับการชักได้ถูกต้อง 93.74% โดยมีค่าเอฟหนึ่ง (F_1) เท่ากับ 28.61% และอัตราการตรวจจับพลาด 1.78 ครั้งต่อชั่วโมง ซึ่งเทียบเท่ากับประสิทธิภาพจากวิธีป่าสุ่มและเหนือกว่าประสิทธิภาพจากการเรียนรู้ของเครื่องวิธีอื่นที่ใช้เวกเตอร์คุณลักษณะที่มีนัยสำคัญ การประมวลผลภายหลังทั้งสองวิธีสามารถเพิ่ม สมรรถนะการตรวจจับได้อย่างมีประสิทธิภาพ เมื่อเราใช้โครงข่ายประสาทเทียมร่วมกับสกอร์เน็ต เราพบว่าสมรรถนะที่ดีที่สุดวัดโดยค่าเอฟหนึ่งเท่ากับ 70.15% มีอัตราการตรวจจับที่ถูกต้องเท่ากับ 91.96% และอัตราการตรวจจับพลาดลดลงเหลือ 0.09 ครั้งต่อชั่วโมง นอกจากนี้ ฟังก์ชันสูญเสียแบบไดซ์เชิงลอการิทึมสามารถทำให้ประสิทธิภาพของการตรวจจับการชักดีขึ้นกว่าการสูญเสียแบบอื่นในกรณีที่ความผิดพลาดจากการตรวจจับของข้อมูลขาเข้าและขาออกมีขนาดใหญ่ นอกจากนี้ เราได้เสนอตัวชี้วัดเวลาแฝงยังผลที่คำนึงถึงการตรวจจับพลาด ทำให้ผลของการใช้ตัวชี้วัดนี้ไม่เพียงแค่สามารถระบุความล่าช้าของการตรวจจับได้เท่านั้น แต่ยังสามารถแยกแยะการกระจายตัวของความล่าช้าได้อีกด้วย

ภาควิชา	วิศวกรรมไฟฟ้า	**ลายมือชื่อนิสิต**	.....
สาขาวิชา	วิศวกรรมไฟฟ้า	**ลายมือชื่อ อ.ที่ปรึกษาหลัก**	.....
ปีการศึกษา	2563	**ลายมือชื่อ อ.ที่ปรึกษาร่วม**	.....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม	.....

6071432821: MAJOR ELECTRICAL ENGINEERING
KEYWORDS: SEIZURE ONSET AND OFFSET DETECTION / EEG / CNN / SCORENET / EL-INDEX
POOMIPAT BOONYAKITANONT : AUTOMATIC DETECTION OF EPILEPTIC SEIZURE ONSET AND OFFSET IN EEG.
ADVISOR : ASSOC. PROF. JITKOMUT SONGSIRI, Ph.D.,
CO-ADVISOR : ASSIST. PROF. APIWAT LEK-UTHAI, Ph.D.,
CO-ADVISOR : ASSIST. PROF. KRISNACHAI CHOMTHO, M.D., 89 pp.

This dissertation presents an automatic detection method of seizure onset and offset in scalp EEGs to reduce the required efforts from experts. In this work, a framework of the seizure onset and offset detection consists of epoch-based seizure detection and seizure onset and offset determination. A deep convolutional neural network (CNN) is designed without requiring any hand-crafted features to detect seizure episodes in the EEGs. Two post-processing techniques called a counting-based method and ScoreNet are proposed to indicate the beginning and ending of the detected episodes. The criterion in the counting-based method is heuristically designed from a clinical fact of continuous changes of EEG patterns during epilepsy, and ScoreNet is a neural network-based model which generalizes the counting-based method by adding more flexibility. Moreover, a log-dice loss is presented as an objective function for training ScoreNet to handle a class imbalance problem. A public database was used in this work to verify the proposed method in a patient-specific scheme. As a result, overall GDR of 93.74%, F_1 of 28.61%, and FPR/h of 1.78 obtained from the proposed CNN were comparable to those from a random forest and superior to those from other machine learning methods with significant features. The proposed post-processing techniques could potentially improve seizure detection performances, and the best F_1 of 70.15% was accomplished by a combination of CNN and ScoreNet, and GDR and FPR/h obtained from this combination were 91.96% and 0.09, respectively. Furthermore, the log-dice loss could better enhance the detection performances than other losses in case of large classification errors between inputs and outputs. In addition, we propose an effective latency index (EL-index) as a metric which includes undetected events into account. This index not only provides a meaning of detection delays but also distinguishes between different distributions of the delays.

Department:	Electrical Engineering	**Student's Signature**	.....
Field of Study:	Electrical Engineering	**Advisor's Signature**	.....
Academic Year:	2020	**Co-advisor's signature**	.....
		Co-advisor's signature	.....

Acknowledgements

First of all, I would like to express my sincere appreciation to my advisors, Associate Professor Jitkomut Songsiri, Assistant Professor Apiwat Lek-uthai, and Assistant Professor Krisnachai Chomtho, for continuously and kindly supporting my Ph.D. work, for strengthening me, and for encouraging me to continue my study.

Besides my advisors, I would like to thank my dissertation committees: Associate Professor Supatana Auethavekiat, Associate Professor Chedsada Chinrungrueng, Associate Professor Setha Pan-ngum, and Dr. Pasin Isarasena na Ayudhya for their precious time, their efforts, and their useful comments. Thanks to Associate Professor Manisa Pipattanasomporn for guidance of English writing.

Thanks to Chulalongkorn University Graduate School for financial supports: The 100^{th} Anniversary Chulalongkorn University for Doctoral Scholarship and The 90^{th} Anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund) for my Ph.D. life.

I also would love to thank CSRL staffs: Professor David Banjerdpongchai, Associate Professor Watcharapong Khovidhungij, Assistant Projessor Suchin Arunsawatwong, and Assistant Professor Manop Wongsaisuwan, CSRL colleagues: Tadchanon Chuman, Pham Van Tuynh, Chanthawit Anuntasethakul and Parinthorn Manomaisaowapak, and CSRL alumni: Anupon Pruttiakaravanich, Suppachai Suksamosorn, Nattaporn Plub-in, Surawit Laohanan, Kebsiri Manusilp, Jeerapat Jitnuant, Tanwalai Panapongpakorn, Ruipart Kruakaew, Petchakrit Pinyopawasutthi, Ranyaphat Hongpipatsak, Pongsorn Keadtipod, and Nattapol Teachapangam for marvelous experiences, sleepless nights, useful and fun discussions, encouragements, and parking lots.

Last but not least, I would like to give the deepest thank to my family for always physically and emotionally supporting me and for motivating me throughout my life.

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

Chapter I

INTRODUCTION

This chapter presents an introduction of this dissertation consisting of three sections. Section 1.1 provides motivation and importance of this work, and an overview of this dissertation is given in Section 1.2. In Section 1.3, problem statements which contain seizure detection and seizure onset and offset determination are explained.

1.1 Motivation

Defined by the International League Against Epilepsy, an epileptic seizure is a transitory occurrence of a symptom due to an abnormal excessive or synchronous neuronal activity in the brain [1]. It was reported that 65 million people of all ages are affected by the epilepsy [2]. Consequences of epilepsy are largely dependent on types of seizures and areas that the seizures appear. For instance, a tonic-clonic seizure can initiate from one side or both sides of the brain, and people affected by the tonic-clonic seizure have uncontrollable, stiffening, and jerking muscle [3]. An absence seizure which is a generalized onset seizure affects patient's awareness. The absence seizure usually has effects on the patient in a short period, usually less than 10 seconds, but there are also absence seizures that last longer [4]. Due to the impacts of epileptic seizures, which can lead to neuronal and physical injuries, patients with recurrent or prolonged seizures should be reviewed by neurologists for a prompt diagnosis and a treatment. Neurologists usually monitor the patients with continuous video-EEG monitoring [5; 6] for those having refractory status epilepticus that are unresponsive to therapy. This is a combination of electroencephalography (EEG) and video, recorded simultaneously to observe brain activities in relation with a clinical change. Nevertheless, this task is still a time-consuming process for the neurologists to continuously review long EEGs, and fatigue from examining the records for too long can lead to inconsistent diagnosis. In addition, it is also challenging to automatically indicate the seizure occurrences in the long EEG records because of the rarity, unpredictability, and diversity of the epileptic seizures. Therefore, automated epileptic seizure detection using EEG signals has been developed to facilitate the analyses of long-term monitoring.

Seizure onset detection has an important role in situations that require immediate treatments, especially in cases when patients do not respond to the medication. There are two types of seizure onset that can be inspected from the scalp EEG signals regarding to the spatial distribution of the seizure activity. When a seizure which originates at some point rapidly spreads the whole networks, causing EEG changes apparently on the whole brain, it is called a generalized-onset seizure. In contrast, a seizure is focal-onset when originating within networks limited to one particular area, making the changes in EEG restricted in a particular brain region [7; 8]. An immediate alarm in a seizure detection system is needed for a patient who requires a treatment after the seizure starts. A detection delay from the actual seizure onset possibly causes late therapy and wrong localization of an epileptogenic focus [9]. For instance, a responsive neurostimulation system is a device that is implanted in the brain to observe and stimulate brain activities [10; 11]. This device releases electricity to reduce an impact of seizure after the seizure initiates, so the detection delay of the seizure onset plausibly makes the treatment less effective. Moreover, the onset of secondarily generalized seizures which is focal onset seizures that quickly spreads to the whole brain is hardly localized. Hence, a nearly correct indication of the seizure onset is needed for a proper treatment.

Moreover, detecting seizure offset is also important to neurologists for a proper diagnosis. Seizure offset recognition can help reduce the side effects in postictal states by a prompt treatment [12]. Providing

a period of an epileptic seizure in an EEG record to neurologists instead of only the occurrence of the seizure can better assist the neurologists to consequently analyze and adjust antiepileptic drug therapy properly [13]. For instance, it is highly possible that the seizure will become status epilepticus if it lasts longer than five minutes without prompt therapy. In this case, lack of treatment can considerably damage the human brain. However, it is not easy to indicate the seizure offset following the seizure activity. There are several possibilities of transitions from the seizure activities to their terminations; the seizure offset cannot be directly observed from the channels where the seizure initiates [14]; it is possible that a focal-onset seizure is still localized or spreads to the whole brain.

From the needs and importance of the automatic detection of epileptic seizures and the starting and ending points, this work mainly concentrates on detecting the seizure events and determining their onsets and offsets. We divide the whole project into two main steps: epoch-based classification and seizure onset and offset detection. The epoch-based classification is to classify EEG epochs segmented from long EEG signals into classes, and the seizure onset and offset detection adopts the epoch-based results to improve the classification performance and indicate the seizure onsets and offsets.

1.2 Dissertation overview

An overview of this dissertation contains the objective, scopes, methodology, and benefits and outcomes of this dissertation. This dissertation is organized as follows. Chapter 2 reviews backgrounds of this dissertation which includes montages and characteristics of EEG. Since this work focuses on using machine learning techniques, empirical risk minimization, which is the most fundamental learning concept in machine learning, descriptions of widely used classification methods, and cost functions are also explained in Chapter 2. Studies related to the detection of seizure events, onsets and offsets are discussed in Chapter 3. Moreover, Chapter 4 describes models for predicting seizure episodes and indicating the onsets and offsets. This chapter also includes a proposed classification loss function, a feature selection method, and a metric of detection delay. Chapter 5 explains an EEG database, evaluation metrics, and a validation scheme. Primary experiments consisting of feature analysis, seizure detection, and seizure onset and offset determination are clarified in Chapter 6. Finally, Chapter 7 summarizes the conclusion, limitations, and future work of this dissertation. Since some details in Chapters 3, 4 and 6 were already published [15; 16; 17; 18], these contents in those chapters are partly taken from these references.

Objectives

The aim of this work is to provide an offline detection method of seizure activities and an identification approach of their starting and ending time points in multi-channel scalp EEG signals to help neurologists label EEG records that have been collected.

Scopes of work

- This work considers a patient-specific detection scheme to control demographics of case studies. Data for training and testing are collected from the same patient, so diversity of patient profiles do not affect the detection performance. We will show preliminary results in Appendix C.3 that, even if we applied the patient-specific scheme, F_1 scores of all patients still had a wide range. Moreover, as reported in Section 6.2, it was impossible to select parameters of a radial basis function for a support vector machine that could detect at least one seizure event from every patient. It is evident that tuning appropriate hyperparameters of a patient non-specific seizure detector is possibly unachievable; thus, we do not aim to develop a single detector for all patients.

- Multi-channel scalp EEG signals are used to detect seizure activities, not intracranial EEG signals, and no other modality of biomedical signal and imaging is included. An annotation of each record must contain the beginning and ending time points of each seizure event. All training and test data must be acquired from the same montage, and the data are collected from an online open source. In addition, there must be at least seizure event in training data to learn the proposed model.
- The training and testing stages are conducted offline. All data must have been readily available before all experiment are conducted. Types of seizures are not necessarily specified in a data set of interest, as we do not discriminate seizure types in this work.
- Comparisons of existing methods are demonstrated. These also include the verification of the proposed seizure onset and offset identification model using different epoch-based seizure classification approaches.

Methodology

This section explains the study plan depicted in Table 1.1 and the methodology of this work by items as follows.

Table 1.1: Study plan.

Items	Semester							
	1	2	3	4	5	6	7	8
Review literature	■	■	■					
Collect online data		■	■					
Write and submit review article			■	■				
Propose seizure onset and offset detection method				■	■			
Prepare proposal examination					■	■	■	
Analyze the seizure onset and offset detection method							■	
Submit technical article							■	■
Conclude the dissertation and prepare the examination								

- Review literature of EEG data, pre-processing, feature extraction, classification, and process of determining the seizure onset and offset in EEG signals.
- Collect data from several subjects where each subject has many records. There must be at least one record of each subject containing at least one seizure activity.
- Propose a method to detect seizure episodes based on EEG epochs using a machine learning tool, and present a technique to indicate the seizure onset and offset.
- Train an epoch-based classifier on EEG segments where the training set must contain at least one seizure event and verify results of the classification on a test set collected from the same patient.
- Apply the proposed onset and offset detection model to the classification results of EEG epochs to determine starting and ending points of the detected seizure episodes. Compare the results of the onset and offset detection and the classification results in a fair setting.
- Conclude the detection performances, limitations, and future work.

Benefits and outcomes

Benefits. Our method provides a pre-annotation of seizures for the neurologists to further analyze seizure characteristics. With our proposed scheme, less efforts than usual from neurologists are required to review the continuous EEG, and a little background knowledge about epilepsy is needed. Our scheme requires a seizure-annotated EEG record containing at least one seizure event for customizing a model for a particular patient. This customized model can then be used to automatically mark expected seizures on unseen EEG records collected from the same patient, so the neurologists have to provide only the annotation of the training set. Moreover, no other modalities, *e.g.*, electrocardiogram (ECG), electromyogram (EMG), are included; no other equipment is required to collect the data.

Outcomes. First, we provide an automatic detection method of epileptic seizures and their seizure onsets and offsets using multi-channel EEG signals. The algorithm associated with the proposed method is also given to train the model using multi-channel scalp EEG signals. Moreover, comparisons between our proposed method and other approaches evaluated from the same set of data are presented in terms of seizure detection and seizure onset and offset determination. In particular, these comparisons include evaluations of the proposed seizure onset and offset detector when other epoch-based seizure classification methods are applied.

1.3 Problem statements

This research aims to detect seizure events in long scalp EEGs and to determine the onsets and offsets of the seizures. Consider a process of detecting the seizure onset and offset in a long EEG signal shown in Figure 1.1. The problem of detecting the onset and offset consists of divided into two sequential problems: finding epileptic seizure episodes in multi-channel EEG signals and indicating the beginning and ending time points of the epileptic seizure activity. In the seizure detection process, each segmented EEG from a long EEG signal is individually used to indicate a seizure occurrence. When all EEG epochs from the long EEG signal are classified using the seizure detection algorithm, the outputs are assembled into the sequence of seizure predictions. Subsequently, the prediction sequence is fed to the onset and offset determination process to indicate the beginning and termination of each individual seizure in the long EEG signal.

Seizure detection

The goal of the seizure classification is to classify EEG epochs into a correct class. Consider the epoch-based seizure classification shown in Figure 1.1. A multi-channel EEG epoch segmented from a long multi-channel EEG signal is considered as a sample. A seizure classifier then receives EEG information as the input and returns either the probability or the decision of a seizure occurrence of epochs as the output:

$$z_i = h(x_i) \tag{1.1}$$

where h is a function representing the classifier, and x_i and z_i are the input and the output of h at the epoch i, respectively. In this case, h can be any classifier including logistic regression, SVM, random forest, and CNN. Depending on the classification method, we can use extracted features or raw EEG data as the input. Similarly, z is either a seizure probability, *e.g.*, $z_i \in [0, 1]$, or a predicted value, *e.g.*, $z_i \in \{0, 1\}$.

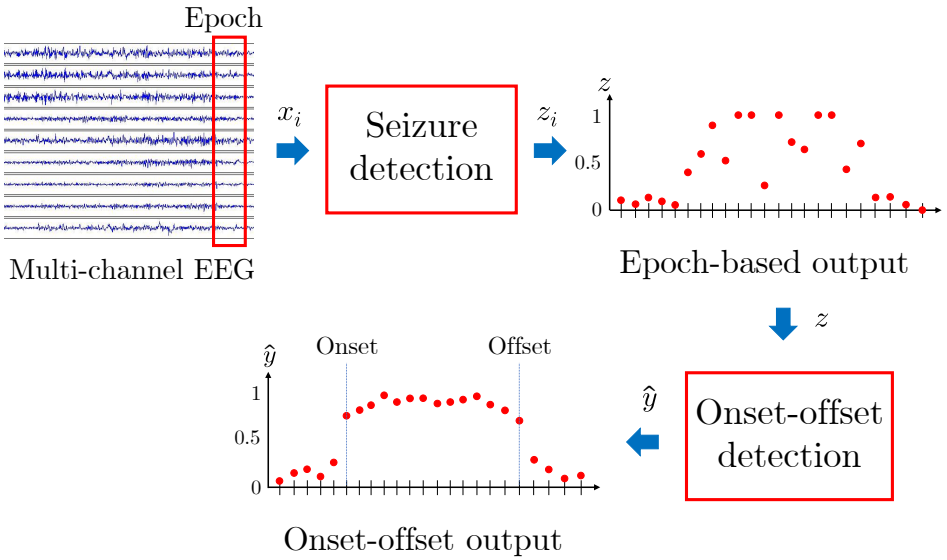


Figure 1.1: Scheme of the problem containing two statements: epoch-based seizure detection and onset-offset detection.

Seizure onset and offset determination

Seizure onset and offset determination is the process of implying the beginning and the ending in a long EEG signal from the sequence of seizure predictions. Let $z = (z_1, z_2, \ldots, z_N)$ be the sequence of the epoch-based seizure detection output, and $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$ be the output sequence of the onset and offset determination, respectively, where $\hat{y}_i$ is the probability that the epoch i is indicated as a seizure, and N is the number of epochs in the long EEG signal. The goal is to find a function $g : [0, 1]^N \to [0, 1]^N$ to modify z so that the output $\hat{y}$ is more clinically realistic for indicating the seizure onset and offset as shown in Figure 1.1:

$$\hat{y} = g(z). \tag{1.2}$$

The seizure onset is determined by the first index k of a predicted group of seizure-detected epochs containing $\hat{y}_k \geq 0.5$. Similarly, the index k implies the seizure offset when k is the last index of the predicted group.

Chapter II

BACKGROUND

In this chapter, required backgrounds are explained to comprehend contents in this dissertation. This chapter is divided into four sections. In Section 2.1, characteristics and montages of EEGs are demonstrated as basic knowledge for designing a classification method of seizures and a detection approach of their starting and ending points. Section 2.2 describes empirical risk minimization to comprehend a framework of machine learning. Next, conventional and existing classification methods that have been widely used are explained in Section 2.3, and details of cost functions in classification problem are finally provided in Section 2.4.

2.1 Electroencephalography (EEG)

EEG and montages

Electroencephalography (EEG) is a technique of recording and studying electric potentials involved with the brain's electrical activities [4]. The studies of the electrical activities in the brain have used EEG records for diagnosing diseases in neuroscience, for example, epilepsy, brain tumors, head injury, and sleep disorders. There are two types of EEGs, scalp and intracranial EEGs, depending on where signals are observed. The scalp EEG signals are recorded by placing small disks called electrodes in different positions on the scalp surface with liquid gel. For the intracranial EEG (iEEG), or so-called electrocorticogram (ECoG), the subdural electrodes are implanted directly in the brain during the surgery to measure the electrical signals directly from the cerebral cortex.

Locations of electrodes on the scalp are critical because the measured signals spatially vary on the position of the scalp; difficulties and mistakes of interpreting brain conditions exist when electrode locations are incorrect. One of the standard placements of electrodes is the international 10-20 electrode system. As shown in Figure 2.1, electrodes are placed with 10% or 20% of actual distances between adjacent electrodes in all three directions. The reference points of the system are nasion, the depressed area between the eyes, and inion, the prominent bone locating on the middle line of the skull. Each location is assigned by a letter to specify a lobe and by a number to specify the location of each lobe. The letters F, T, C, P and O are used in the positions of Frontal, Temporal, Central, Parietal and Occipital lobes, respectively, and the midline of the brain is indicated by Z. Odd numbers and even numbers refer electrodes on the left and right hemispheres, respectively.

EEG recordings are monitored in the various way according to montages, the placements of the electrodes. Two popularized montages that are currently used are bipolar and referential montages. In the bipolar montage, a pair of adjacent electrodes are inputs to a differential amplifier resulting a waveform of each channel displayed on the monitor. On the other hand, the referential montage is a montage that the output of each channel is the voltage difference between a certain electrode and a common reference electrode. Generally, there is no standard position for the reference; however, the linked ears, referring to the positions A_1 and A_2, and midline positions are often preferred as a reference. When the common reference is a voltage averaged over the brain, the montage is called an average reference montage.

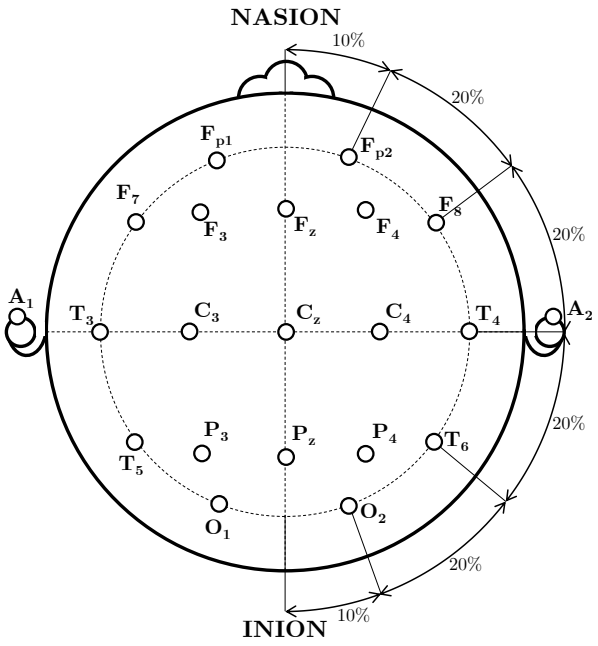


Figure 2.1: Illustration of the international 10-20 system from the top view of the head.

EEG characteristics

There are four main rhythms of the normal EEG, namely, alpha, beta, theta, and delta, that need to be primarily described [4]. Neurologists normally use the knowledge of normal EEG patterns to visually identify the epileptic seizures in the long EEG signals. A rhythm called Alpha rhythm is considered as the principal background of the normal EEG occurring in a frequency range of 8 - 13 Hz. This rhythm is predominantly observed when the patient is relaxed in a waking state with closed eyes. It is usually maximum in the occipital area and spreads to the adjacent regions, *e.g.*, parietal and temporal regions. Patterns of the beta rhythm (14 - 30 Hz or higher) appears with longer duration than muscle action potentials. Theta rhythm is defined as an activity in a frequency band of 4 - 7 Hz. It is typically dominant in the midline and the temporal region. This rhythm indicates a waking and drowsiness state and should be symmetrically diffused. If the theta activity appears only in one area or one hemisphere, this may refer to structural disease. Delta rhythm is a slow wave that its frequency distributes in 0.5 - 4 Hz. This wave usually has high amplitudes and can be observed during a deep sleep stage. Nevertheless, this wave is also prominent to implications of cerebral dysfunction if seen during wakefulness in adults [4].

On the other hand, epileptiform patterns in EEG signals are abnormal patterns used to indicate epileptic seizures in the long EEG signals. By definition, the epileptiform patterns are *spikes* and *spike-wave complexes*; however, other abnormal patterns such as *sharp waves* are also important to the detection of the epileptic seizures [19]. The definition of the *spike* is an abrupt change of temporal potential from the background where its decline slope is lower than that of the incline. The spike duration is less than 70 milliseconds, the length of the sharp wave is 70 - 200 milliseconds, the slow wave lasts more than 200 milliseconds. The *spike-slow-wave complex*, also commonly called a *spike wave*, contains the spike and a following slow wave with relatively high amplitudes. In addition, the *sharp wave* is practically essential in determining the epileptic seizure even though it is not demonstrated as an epileptic pattern. The sharp wave represents a wave with a frequency of 5 - 12.5 Hz. A sequence of spike, sharp, and spike-slow wave is referred to ictal patterns of EEG when epileptic seizures occur. By the morphology of these three patterns, *i.e.*, spikes, spike-slow waves, and sharp waves, changes in amplitudes, frequencies, and rhythms continuously established relative to the background [20]. During seizure activities,

amplitudes of EEG signal during epileptic seizure activities tend to be higher than those of normal periods, a frequency shift appears when brain activities transit from normal events, *e.g.*, drowsiness, eye blink, to the seizure activities, and rhythms or patterns in EEG signals change from normal patterns to specific patterns. However, some change seems to be an occurrence of epileptic seizures even though this change is referred to an artifact. For instance, EEG signals interfered by main electricity have evolution of amplitudes from low to high and then still maintain the amplitudes at this level for a course of time. Moreover, periodic epileptiform discharges (PED) are also uncommon EEG characteristics similar to seizure activities but determined as non-seizure activities. This makes seizure detection challenging in discriminating the ictal patterns from interictal EEG signals.

2.2 Empirical risk minimization

Empirical risk minimization is a statistical learning principle in machine learning used to theoretically indicate bounds of performances of learning algorithms [21; 22]. Suppose that $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$ is a space of pairs (x_i, y_i) where $\mathcal{X}$ and $\mathcal{Y}$ are spaces of all inputs and outputs, respectively. Formally, there is a joint probability distribution $f_{xy}(x, y)$ over $\mathcal{D}$, and (x_i, y_i) is drawn from the distribution f_{xy}. In supervised machine learning problems, there exists an actual function that maps every input sample $x_i \in \mathcal{X}$ to its label $y_i \in \mathcal{Y}$. The major goal is to find a mapping function called a hypothesis or a learner h in a hypothesis space $\mathcal{H}$ that approximately behaves like the actual function: $h(x_i) \approx y_i, \forall (x_i, y_i) \in \mathcal{D}$.

A loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbf{R}^+ \cup \{0\}$ is a non-negative-valued function that quantifies how accurate the classifier is from a difference between an output and a label. For instance, a 0-1 loss function, which disregards a correct classification but absolutely focuses on an incorrect result, is defined as

$$L(h(x_i), y_i) = \begin{cases} 0, & h(x_i) = y_i, \\ 1, & \text{otherwise.} \end{cases} \tag{2.1}$$

The *true error*, also called the *expected risk* and the *Bayes risk*, is defined as the expected value of the loss function over the distribution f_{xy} to measure the overall error of the results from the classifiers:

$$R_{\text{true}}(h) = \mathbf{E}[L(h(x), y)]. \tag{2.2}$$

Since $\mathcal{Y}$ contains only discrete elements, the *true error* becomes

$$R_{\text{true}}(h) = \int \sum_{y \in \mathcal{Y}} f_{xy}(x, y) L(h(x), y) dx. \tag{2.3}$$

The main problem is to find the optimal learner h^* in the hypothesis space $\mathcal{H}$ such that it minimizes $R_{\text{true}}(h)$:

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} R_{\text{true}}(h). \tag{2.4}$$

The optimal hypothesis h^* is formally called the *Bayes optimal classifier*, and the minimum error $R_{\text{true}}(h^*)$ is named as the *Bayes error rate*.

However, $R_{\text{true}}(h)$ cannot be directly obtained from (2.3), and it cannot be minimized since $f_{xy}(x, y)$ is practically unknown. Hence, the *empirical error* using data in $\mathcal{D}$ is employed as the estimation of $R_{\text{true}}(h)$:

$$R_{\text{emp}}(h) = \sum_{(x_i, y_i) \in \mathcal{D}} P(x_i, y_i) L(h(x_i), y_i), \tag{2.5}$$

where $P(x_i, y_i)$ is the hypothetical joint probability. Nevertheless, the joint probability is also generally unknown. Therefore, it is assumed to be $1/|\mathcal{D}|$ where $|\mathcal{D}|$ is the number of samples in set $\mathcal{D}$:

$$R_{\text{emp}}(h) = \frac{1}{|D|} \sum_{(x_i, y_i) \in \mathcal{D}} L(h(x_i), y_i). \tag{2.6}$$

The optimal learner h^* is, therefore, obtained by minimizing $R_{\text{emp}}(h)$:

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, R_{\text{emp}}(h). \tag{2.7}$$

This approximation in (2.6) is known to be an average of the loss L on the set $\mathcal{D}$. In a machine learning framework, we want to find the optimal learner over the hypothesis space, *e.g.*, a space of linear classifiers or polynomial classifiers, that can properly classify samples in a training set. Therefore, the empirical risk minimization is the main concept in the classification problem to obtain the optimal learner.

2.3 Classification methods

The aim of an epoch-based seizure detection method is to independently classify EEG epochs, which are segmented from multi-channel long EEG signals, into an appropriate class. In this section, we describe details of classifiers, namely, logistic regression, support vector machine (SVM), decision tree, random forest, and convolutional neural network (CNN), that have been commonly used in the epileptic seizure detection. In this section, we denote $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$ a set of inputs of a classification model and $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ a set of labels, and a classifier is presented by h. We consider only a two-class classification problem, so $y_i \in \{0, 1\}$, unless there is an additional definition of the target stated. In what follows, we define a loss which indicates an overall classification error to be $\mathcal{L}$ and a loss from each misclassified sample to be L, *i.e.*, $\mathcal{L}(\mathbf{x}, \mathbf{y}) = (1/N) \sum_i L(h(x_i), y_i)$.

Logistic regression

Logistic regressions are supervised machine learning models that use the logistic function to classify an input into a class by the value of probability [23]. The probability of $y_i = 1$ is modelled as

$$P(y_i = 1 | x_i) = h(x_i) = \frac{1}{1 + e^{-(w^T x_i + b)}}$$

where w and b are model parameters. Typically, a standard approach to find the model parameters in a classification problem is to minimize the negative log-likelihood function of the Bernoulli distribution. The cross-entropy, derived from the negative log-likelihood function, that is commonly used is given by

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log h(x_i) + (1 - y_i) \log(1 - h(x_i))].$$

The model parameters are finally obtained by minimizing the cost function $\mathcal{L}(\mathbf{x}, \mathbf{y})$ over w and b.

In addition, a regularization technique such as l_1 or l_2 regularization can be added to the likelihood function for more generalization of the model. For instance, the l_2 regularization is applied to reduce the weight magnitudes so that the model is not sensitive to data variation, and the cost function is given by

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log h(x_i) + (1 - y_i) \log(1 - h(x_i))] + (\gamma/2) \|w\|_2^2.$$

where γ is a regularization parameter. On the other hand, the l_1 regularization is used to promote many zeros in the model parameters by adding the l_1-norm of the model weights to the likelihood function:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = -\frac{1}{N} \sum_{i=1}^{m} [y_i \log h(x_i) + (1 - y_i) \log(1 - h(x_i))] + \gamma \|w\|_1,$$

Support vector machine

Support vector machines (SVMs) are classification models that separate data into class using a fitted decision boundary given by $w^T\phi(x) + b$ where w and b are model parameters, and $\phi(x)$ is a feature mapping of an input x [23]. The mapping function ϕ transforms the data so that the transformed data between classes are linearly separable as demonstrated in Figure 2.2, and the mapping function is simplified to x when the data can be originally separated by a linear decision boundary. Given the target $y_i \in \{1, -1\}$, the decision $h(x)$ of SVM is in the form of

$$h(x_i) = \textbf{sign}(w^T\phi(x_i) + b), \tag{2.8}$$

where $\textbf{sign}(x)$ is the signum function. The goal is to find the parameters w and b such that the margin between the decision boundary and the nearest point is maximized, and that the number of incorrectly classified samples is minimized. These parameters can be obtained by solving the optimization problem

$$\begin{array}{ll} \underset{w,b}{\text{minimize}} & (1/2)\|w\|_2^2 + C\sum_{i=1}^{m}\zeta_i \\ \text{subject to} & y_i(w^T\phi(x_i) + b) \geq 1 - \zeta_i, \\ & \zeta_i \geq 0, \quad i = 1, \ldots, m, \end{array} \tag{2.9}$$

where $\zeta_1, \zeta_2, \ldots, \zeta_m$ are slack variables allowing some data to be incorrectly classified and C is a regularization parameter that controls the penalty of the slack variables. It is well-known that the optimal parameters w and b from solving (2.9) are given by

$$w = \sum_{i=1}^{m} a_i y_i \phi(x_i) \quad \text{and} \quad b = \frac{1}{N_{\mathcal{M}}} \sum_{i \in \mathcal{M}} \left(y_i - \sum_{j \in S} a_j y_j K(x_i, x_j) \right),$$

where $a_1, \ldots, a_m$ are Lagrange multipliers, $K(x, y) = \phi(x)^T\phi(y)$ is called a kernel function, S is the set of indices of support vectors, $\mathcal{M}$ presents the set of indices of data that have $0 < a_i < C$, and $N_{\mathcal{M}}$ is the size of the set $\mathcal{M}$. Substituting the obtained w in (2.8) yields

$$h(x_i) = \textbf{sign}\left(\sum_{j \in S} a_j y_j K(x_i, x_j) + b \right). \tag{2.10}$$

A kernel function $K(x, y)$ is usually used when the input feature in a classification problem are not linearly separable. Common kernels, *e.g.*, linear kernel, polynomial kernel, and radial basis function (RBF) kernel, are summarized in Table 2.1 where c and d are parameters to be selected to construct the polynomial mapping $\phi(x)$, and σ is a bandwidth parameter. The linear kernel is normally used when the data is linearly separable, whereas the polynomial kernel is applied when the data can be classified by a polynomial function. The RBF kernel is generally exploited when the distance of the original features in the same class is small.

Decision tree

Decision trees are tree-based supervised machine learning models that predict a target as illustrated in Figure 2.3 where the root node receives an input, leaf nodes provide outcomes of the model, and internal nodes show decision criteria [22]. The idea of the decision tree is to segment a feature space into regions in which the data are correctly classified by simple rules. For instance, Figure 2.4 shows a feature space which is split into 4 regions by the decision tree using a set of conditions on each variable.

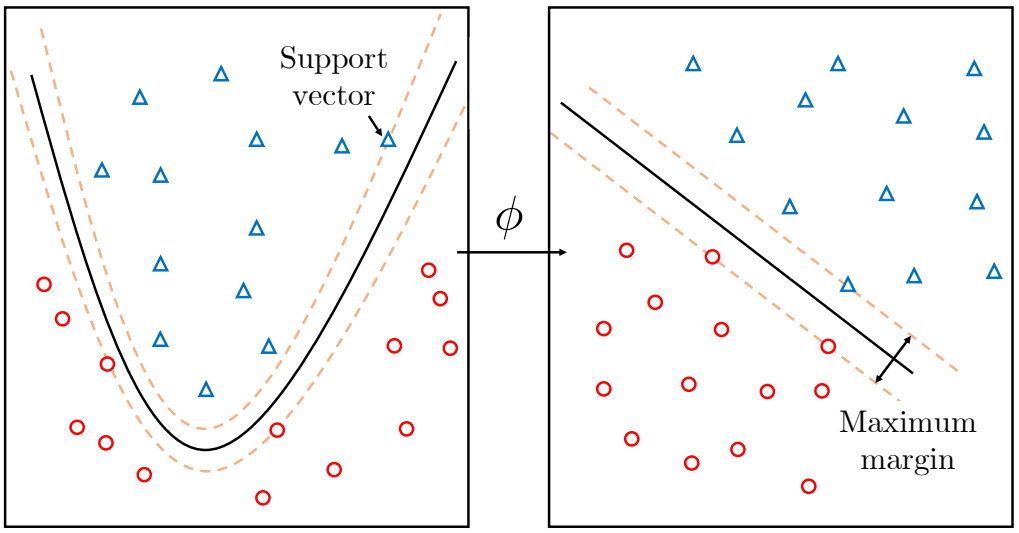


Figure 2.2: Illustration of kernel function. The left side shows the input space, and the right side shows the transformed feature space.

Table 2.1: Commonly used kernels in SVM.

Kernel	**Formula**
Linear	$K(x, y) = x^T y$
Polynomial	$K(x, y) = (c + x^T y)^d$
RBF	$K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$

The model is simple to comprehend and interpret since the decision tree can be clearly visualized. The structure of the decision tree can be more complex when the number of nodes and the depth of the tree increase to fit more complex data. However, with more depth and leaves, the decision tree can be easily overfitting due to small variation in the data. Many small regions containing noises are typically generated by a large tree. For example, region R_4 in Figure 2.4 contains only one sample surrounded by other class samples. In this case, region R_4 should not be created since the sample seems to be corrupted by a noise.

In a classification problem, the goal is to find optimal rules θ in the decision tree such that a decision is correct for a given input. Here, the optimal rules appear to be conditions for generating regions that best separate training data into appropriate classes. To learn the set, suppose that R_m is a region generated from node m, and $\hat{p}_m$ denotes the proportion of class labelled by $y = 1$ in R_m:

$$\hat{p}_m = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = 1),$$

where N_m is the number of samples in R_m, and $I(x)$ is the indicator function. A sample x in node m is classified to the majority class, a class with the highest proportion $\hat{p}_m$:

$$h(x_i) = \begin{cases} 1, & \hat{p}_m \geq 0.5, \\ 0, & \text{otherwise} \end{cases}.$$

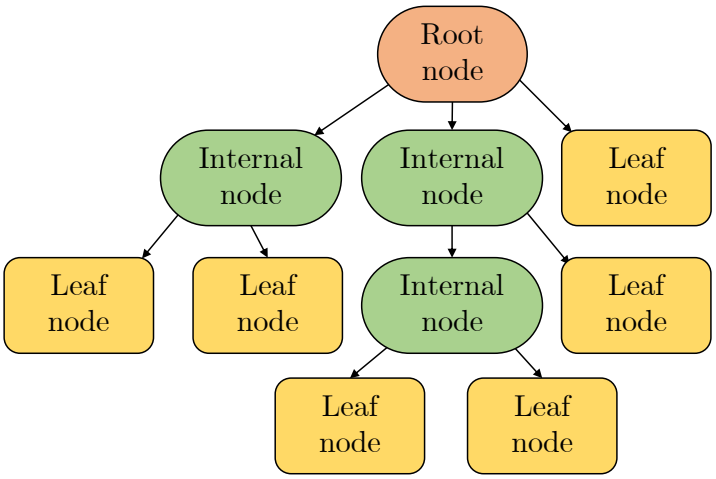


Figure 2.3: Model of decision tree. An arrow shows a flow of data from node to node.

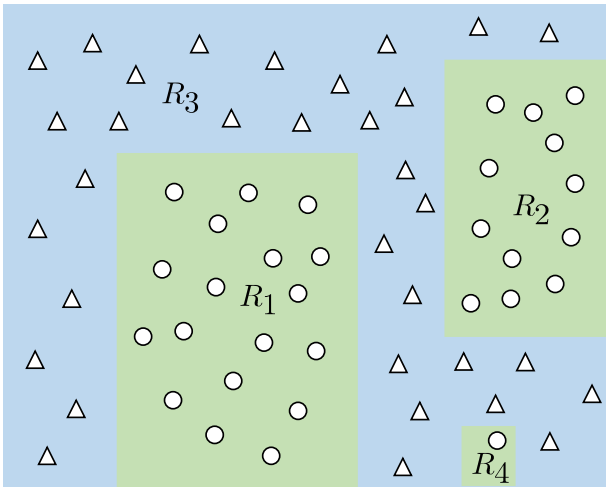


Figure 2.4: Feature space split into regions by decision tree using simple conditions on each variable where regions R_1, R_2, and R_4 are of the same class.

Therefore, the set of the optimal rules is obtained by solving the following optimization problem

$$\underset{\theta}{\text{minimize}} \quad \sum_{m=1}^{M} N_m \sum_{x_i \in R_m} L(h(x_i), y_i) + \alpha M,$$

where M is the number of leaf nodes and α is a regularization parameter controlling the tree size and accuracy of the decision. Three widely used loss functions to measure a classification error are as follows:

- Misclassification rate: $\sum_{x_i \in R_m} L(h(x_i), y_i) = 1 - \max(\hat{p}_m, 1 - \hat{p}_m)$.
- Gini impurity: $\sum_{x_i \in R_m} L(h(x_i), y_i) = 2\hat{p}_m(1 - \hat{p}_m)$.
- Cross-entropy: $\sum_{x_i \in R_m} L(h(x_i), y_i) = -\hat{p}_m \log \hat{p}_m - (1 - \hat{p}_m) \log(1 - \hat{p}_m)$.

It has been shown that the characteristics of these losses are similar [22]. However, the difference should be taken into account to choose a suitable loss. The misclassification rate is less sensitives to perturbation in $\hat{p}_m$ than the others. On the other hand, the Gini impurity and the cross-entropy are numerically stable because they are differentiable. In addition, the computational complexity of the Gini impurity is less complex than that of the cross-entropy since the cross-entropy uses logarithms in its formula; thus, it is faster to calculate the Gini impurity than the cross-entropy.

As mentioned, the complexity of a decision tree is mainly related to the tree structure, and the tree is more complex as the depth and the number of nodes increase. Tree growing begins from the root node, and the node is successively split when there is an impurity produced by the node. This process terminates when all samples are correctly classified by the tree. Nonetheless, the resulted tree is possibly too large and prone to be overfitting since it creates many small regions for a few data. One way to prevent this problem is to set the maximum depth of the tree and the maximum number of leaf nodes. These parameters directly limit the tree size so that the tree is not considerably large. Nevertheless, if these values are low, the tree is probably too small to be applied. Another condition is to assign the minimum number of samples needed to split an internal node. This condition restricts the tree from growing if the region corresponding to the internal node contains only a few samples. In addition, the minimum number of samples in leaf nodes can also be used to avoid creating small regions for a few data. However, the tree growing is hardly proceeded when these minimum numbers are high; thus, the tree becomes too small to be used.

Random forest

Random forests are also tree-based models that aggregate outcomes of multiple decision trees to reduce overfitting in a single tree [24]. The idea is to create several decision trees by randomizing training data independently. The random term here refers to randomization of features and an amount of features used to generate trees to reduce bias from each tree. By averaging outputs of all trees, it has been proved that, with this framework, the random forest is relatively more robust to outliers and disturbances in data than a single decision tree [24]. Moreover, all trees can be parallelly generated, so the random forest is naturally built faster than other ensemble methods such as boosting.

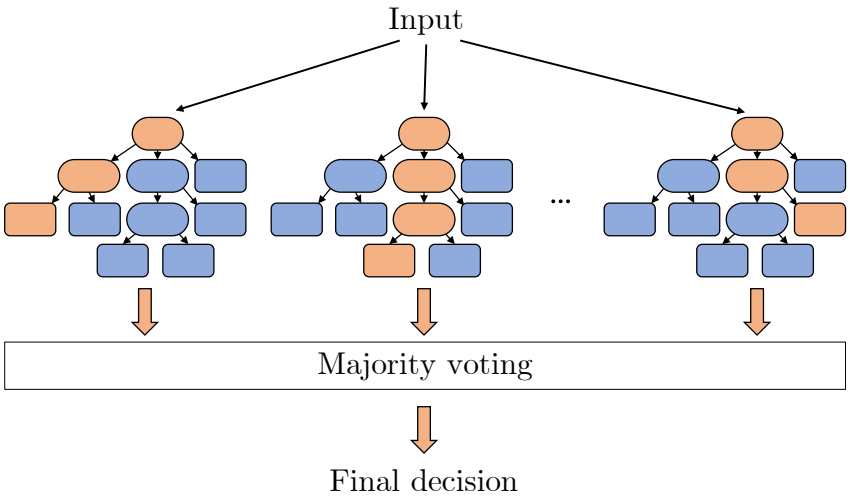


Figure 2.5: Model of random forests. Orange nodes represent flows of data in each tree.

Consider a random forest used in a classification problem shown in Figure 2.5. Let $h_1, \ldots, h_n$ be trees in the random forest. Each tree is independently trained in parallel using a set of random data. To

predict a new input x_i, the probability of $y_i = 1$ can be observed by averaging all outputs of the trees

$$h(x_i) = \frac{1}{n}\sum_{j=1}^{n} h_j(x_i),$$

and the final class is selected from the majority class; when $h(x_i) \geq 0.5$, the class labelled by $y_i = 1$ is chosen in a binary classification problem.

Convolutional neural network (CNN)

CNN is a type of neural networks that has been widely used in various applications: image processing, object detection, face recognition, natural language processing, and video processing [25]. For example, VGG16 is a deep CNN that achieves top-5 accuracy in the ImageNet data set [26]. The CNN is biologically inspired by the idea of animal vision that concentrates on a specific area of an image, called receptive field, instead of focusing on the whole image. The main advantages of this network are that it has spatial invariance property and less computational complexity because of the weight-sharing architecture of convolutional layers [27]. The CNN structure mainly consists of convolutional, activation, pooling, and fully connected layers stacked deeply as demonstrated in Figure 2.6. The computations of the convolutional, activation, and pooling layers are visualized in Figure 2.7. Some regularization technique such as dropout is also added to reduce the effect of an overfitting problem [28], and a batch normalization layer is used to enhance the learning speed [29].

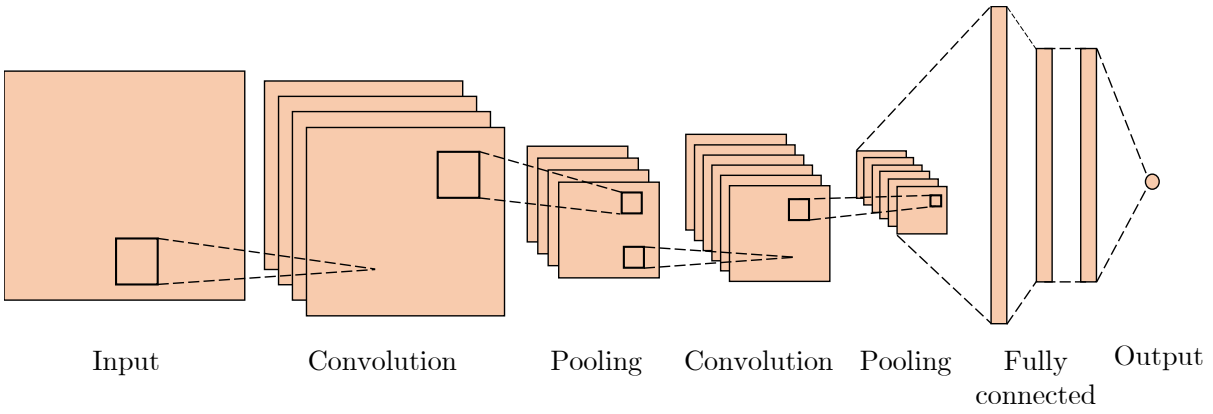


Figure 2.6: Convolutional neural network structure.

The convolutional layer is a layer in which each neuron is locally connected to some area in the previous layer. This layer is mainly designed to extract and collect low-level and high-level features from each layer [27]. The result of each neuron is obtained by multiplying the local input by weights of filters. As shown in Figure 2.7a, the convolutional layer is a result of convolution of the input and the weights. The result can be visually interpreted as a feature map extracted on the receptive field. Hence, to extract many features simultaneously in the same layer, independent filters stacked in depth are used instead of only one filter.

The activation layer visualizes active nodes using an activation function. The output of every node in the previous layer is independently passed to the activation function. Additionally, the activation function can also be physically interpreted as a function that activates and deactivates each neuron in the layer. An example of using activation functions transforming a feature map is illustrated in Figure 2.7b. Common activation functions are listed with their benefits and drawbacks as follows.

- *Identity function* is a function that the output and input are the same:

$$f(z) = z, \quad \frac{d}{dz} f(z) = 1. \tag{2.11}$$

The identity function is put in the output layer when a regression problem is considered. However, it is well-known that the activation function in hidden layers should not be the identity function because if that is the case, then the output is only a linear transformation of the input.

- *Sigmoid function* (σ), or logistic function, is a common activation function used in neural networks. The output of the function is the conditional probability given the input:

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad \frac{d\sigma(z)}{dz} = \sigma(z)(1 - \sigma(z)). \tag{2.12}$$

The advantages of this function are that it is differentiable at every point, bounded, and monotonic, so it is also known as a smooth version of the step function. However, when z is largely positive and negative, the slope of the curve becomes to small, increasing training time; this problem is called a vanishing gradient. The sigmoid function also has a shift bias, causing the network to learn slowly [30].

- *Hyperbolic tangent* (*tanh*) *function* is a function that is similar to the sigmoid function that it is bounded. Unlike the sigmoid function, the output of the tanh function is in the range of $(-1, 1)$:

$$\tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}, \quad \frac{d\tanh(z)}{dz} = 1 - \tanh^2(z). \tag{2.13}$$

The tanh function is used to overcome the shifted bias problem; however, the vanishing gradient problem still occurs.

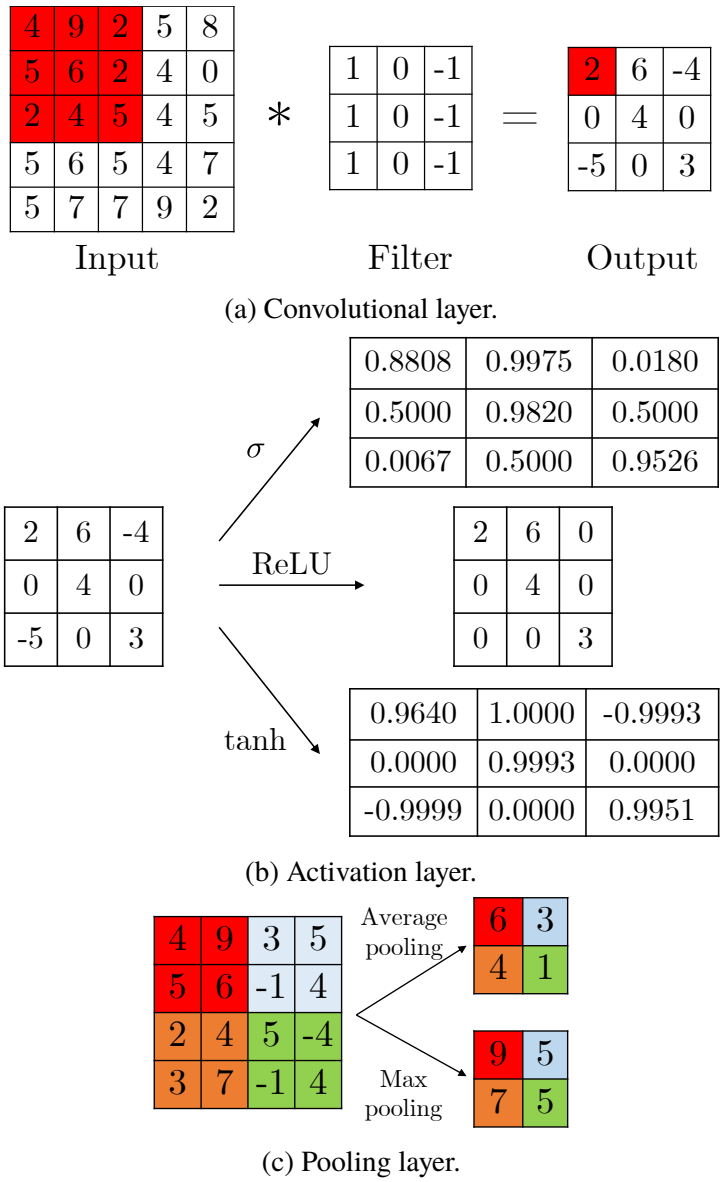
- *Rectified linear unit* (*ReLU*) *function* is a piece-wise linear function that provides zero output when the input is negative, and passes the input to the output when the input is positive:

$$\text{ReLU}(z) = \max(0, z), \quad \frac{d}{dz}\text{ReLU}(z) = \begin{cases} 0, & z < 0, \\ 1, & z > 0. \end{cases} \tag{2.14}$$

The main advantage of using the ReLU function is its computational efficiency for both forward and backward propagation [31]. Moreover, the ReLU function overcomes the vanishing gradient problem when z is large since its derivative is always one. It has also been shown that, in practice, using the ReLU function provides greater convergence performance than using the sigmoid function. However, the function is not differentiable when $z = 0$, and the network learning is prohibited when there are several dead neurons, the neurons that initially give zero outputs always provide zero outputs.

The pooling layer is a layer used extract some appropriate features from the previous layer. When an input is two-dimensional, an image for example, this can be interpreted as performing downsampling along the first and second dimensions of the input. It can be intuitively considered as collecting useful information from the previous layer and filtering out some spatially unnecessary parts. Two common pooling strategies are max pooling and average pooling. As depicted in Figure 2.7c, the max pooling passes the highest value from the receptive field, while the average pooling does average the values in the window.

The batch normalization layer normalizes each input features independently at each mini-batch so that the mean of features is zero and the variance of features closes to one [29]. According to the ability



(a) Convolutional layer.

(b) Activation layer.

(c) Pooling layer.

Figure 2.7: Computation of convolutional, activation, and pooling layers in CNN.

to extract features in each layer, each neuron in the feature map possibly has different mean and variance. Moreover, distributions of the activation outputs are also changed during training since the weights are adapted continuously. This problem is called *Internal Covariate Shift*, and it affects the learning speed. This layer is added to enhance the network to converge faster and prevent the network from the internal covariate shift. Considering a mini-batch $\mathcal{B} = \{x_1, x_2, \ldots, x_k\}$ of the layer inputs, the process of the batch normalization is demonstrated in Algorithm 1 where ϵ is a positive constant preventing numerical instability, and parameters γ and β are to be trained using a backpropagation algorithm [29].

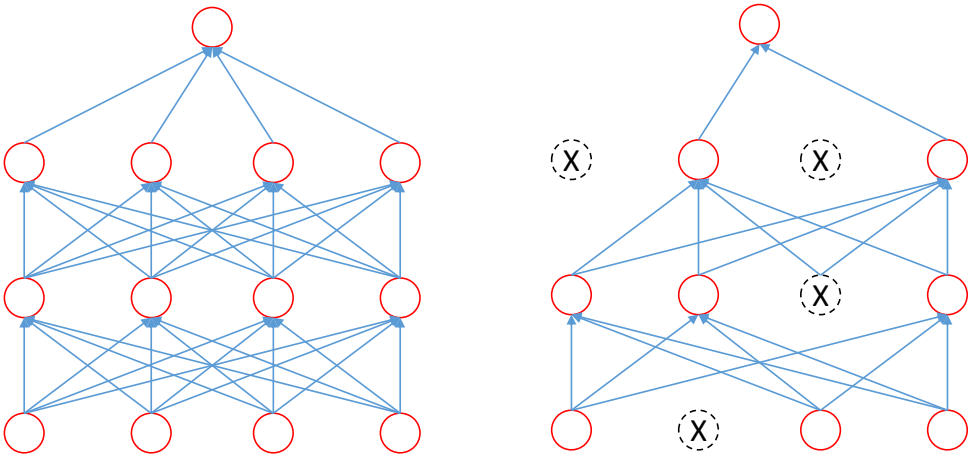
The dropout layer is added to randomly and temporarily removes some neurons in the input layer to prevent overfitting [28]. The idea is to reduce joint adjustments of weights that are suitable for training data. At each iteration, the dropout technique temporarily sets neurons to be inactive with a probability

Algorithm 1: Batch normalization

Input: x over a mini-batch: $\mathcal{B} = \{x_1, x_2, \ldots, x_k\}$
Parameter: γ, β
Output: $\{y_i = \gamma x_i + \beta\}$

1 $\mu_{\mathcal{B}} \leftarrow \frac{1}{k}\sum_{i=1}^{k} x_i$ // mean of mini-batch
2 $\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{k}\sum_{i=1}^{k} (x_i - \mu_{\mathcal{B}})^2$ // variance of mini-batch
3 $\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}}$ // normalization
4 $y_i \leftarrow \gamma \hat{x}_i + \beta$ // scale and shift

p during training as shown in Figure 2.8. The dead neurons in the layer are untrainable, so the weights that need to be trained are only the remaining connections. For testing, all weights are presented, and the weights are multiplied by the probability p to make the node outputs equal to the expected outputs over the distribution used during training. Furthermore, the dropout is also claimed to be superior to other regularization techniques because a neural network with the dropout provides a lower generalization error [28].



(a) Standard neural network. (b) After employing dropout.

Figure 2.8: Dropout in neural network. The neurons with cross signs are temporarily removed from the network.

The fully-connected layer is a layer containing neurons that are all connected to every neuron in the adjacent layers where each connection presents a weight that links two neurons. As demonstrated in Figure 2.6, in deep learning, the fully-connected layer is usually added in the last layer because of its capability of classifying features from the input. Finally, similar to the other machine learning approaches, the model parameters can be obtained by minimizing a cost function, and the most widely used loss function is the cross-entropy.

2.4 Cost functions in classification problem

Given a classifier h, a loss function $\mathcal{L}$ in a classification problem provides an overall similarity measure between actual outputs and predicted outputs in a certain way. Here, three existing cost functions widely used in the classification problem, namely, binary cross entropy, soft-dice loss [32], and squared-dice loss [33] are discussed. Note that both dice losses were mentioned as *soft-dice loss* in their studies; we call them differently to avoid confusion. In a classification problem, a confusion matrix which consists of numbers of true positives, true negatives, false positives, and false positive is considered to evaluate a classification method. We then describe the loss functions explained in this section by these terms for interpretations in this section.

Cross-entropy

The cross entropy is a common loss in classification problem which is derived from the probability mass function of the Bernoulli distribution:

$$\mathcal{L}_{\text{ent}}(\mathbf{x}, \mathbf{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log h(x_i) + (1-y_i)\log\left(1-h(x_i)\right)\right]. \tag{2.15}$$

We observe that $\sum_i (1-y_i)\log(1-h(x_i))$ and $\sum_i y_i \log h(x_i)$ intuitively reflect costs corresponding to false positives and false negatives, respectively. This means that minimizing the cross-entropy is similar to maximizing classification accuracy. However, this loss penalizes all incorrect samples equally; a majority class does provide more contribution than a minority class when the data are highly imbalanced. In this case, the minority class is likely to be ignored; thus, this loss is typically inappropriate to be used [34].

Soft-dice loss

The soft-dice loss ($\mathcal{L}_{\text{softDL}}$) is a function motivated by a dice similarity coefficient (DSC) used to measure a similarity between two sets [32]. DSC is an index for determining an analogy of two set of samples. In a classification problem, DSC, also known as F_1, indicates the overlap between true and predicted values defined by

$$\text{DSC} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{2.16}$$

when TP, FP, and FN stand for numbers of true positives, false positives, and false negatives, respectively. $\mathcal{L}_{\text{softDL}}$ is then defined by

$$\mathcal{L}_{\text{softDL}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{2\sum_{i=1}^{N} y_i h(x_i)}{\sum_{i=1}^{N}\left(y_i + h(x_i)\right)}. \tag{2.17}$$

The second term of RHS in (2.17) is obtained by substituting $\sum_i y_i h(x_i)$, $\sum_i (1-y_i)h(x_i)$, and $\sum_i y_i(1-h(x_i))$ for TP, FP, and FN in DSC, respectively. Therefore, minimizing $\mathcal{L}_{\text{softDL}}$ is similar to maximizing F_1. In addition, DSC does not directly use correctly classified majority samples, *i.e.*, true negatives, so this coefficient has no bias towards the majority class. Thus, when the data is imbalanced, it is an appropriate evaluation metric in the classification task, and the loss yields a suitable property to use in the class-imbalanced problem.

Squared-dice loss

The squared-dice loss ($\mathcal{L}_{\text{sqDL}}$) was originally inspired by DSC for an imbalanced data problem [33]. In contrast to $\mathcal{L}_{\text{softDL}}$, $\mathcal{L}_{\text{sqDL}}$ uses the sum of squares of the actual and predicted labels in the denominator:

$$\mathcal{L}_{\text{sqDL}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{2\sum_{i=1}^{N} y_i h(x_i)}{\sum_{i=1}^{N} \left(y_i^2 + (h(x_i))^2\right)}. \tag{2.18}$$

Similar to the soft-dice loss, the squared-dice loss does not take true negatives into account. In the true negative case, which is $y_i = 0$ and $h(x_i) \approx 0$, $y_i h(x_i)$ and $y_i^2 + (h(x_i))^2$ are all zeros, so the remaining terms in (2.18) are for true positive, false positive, and false negative cases. Therefore, the square-dice loss is also an appropriate loss to determine a classification error in the class-imbalanced problem.

Chapter III

LITERATURE REVIEW

In this chapter, we review previous research related to seizure onset and offset detection using EEGs. In our opinion, to indicate the onset and offset of an epileptic seizure, the seizure is required to be detected first. Hence, this review of literature is primarily divided into two categories: automated epileptic seizure detection and application of seizure onset and offset detection. In Section 3.1, we firstly explain an overall scheme usually used in the automatic epileptic seizure detection. Widely used features and their benefits are then summarized as they are essential to the detection process. Up to now, the CHB-MIT Scalp EEG database [35] is the only data set that meets the scopes of this dissertation, so only seizure detection methods performed on this database are further described and compared. Subsequently, details of previous detection methods of seizure onset and offset are summarized in Section 3.2.

3.1 Automated epileptic seizure detection

Overall, the automatic detection of epileptic seizures normally contains processes of feature extraction, and classification while a signal transformation or decomposition is also commonly but not necessarily used [36; 37; 16]. Artifact or noise rejection has been frequently added at the beginning of the detection process to reduce noises from the signals [38]. In addition, a channel selection technique has been considered in case of multi-channel EEG signals [37], and feature dimension reduction or feature selection algorithms have been taken into account when inputs have a considerably large magnitude [39].

Features are observable quantities used to determine characteristics or properties of events. In a classification problem, the features should be chosen appropriately to be distinguishable between classes. Many features have been employed to discriminate ictal patterns from normal activities in EEG [36; 40; 16]. These features are generally categorized according to the purpose of the work. Some studies employed a group of features according to their meanings and interpretations [41; 42; 43; 44; 45], while others used features according to the domain from which they were extracted [46; 47; 40; 48]. For instance, entropy-based features such as Shannon entropy and approximate entropy were commonly applied to measure the fluctuation of the signal [49; 50; 51; 52]. Using amplitude-related features including nonlinear energy [53] and variance has shown a significant performance of detecting seizure activities with high amplitudes [54; 55; 56; 57]. On the other hand, features are also categorized into time domain, frequency domain, and time-frequency domains regarding signal transformation before the feature extraction. Time-domain features are computed on raw or decomposed signals, intrinsic mode functions (IMFs) from empirical mode decomposition (EMD) for example, in the time domain [58; 52; 59], whereas frequency-domain features are calculated discrete-Fourier transform (DFT) or power spectral density (PSD) coefficients of raw EEG signals [60; 61]. On the other hand, time-frequency-domain attributes are obtained from transformed EEG signals containing both time and frequency information [62; 63; 64; 65; 44]. For instance, coefficients of short-time Fourier transform (STFT) or discrete-wavelet transform (DWT) are used in feature extraction [42; 52]. From literature, we can conclude that statistical parameters, energy and entropies were common features in those three domains to capture information about distributions, amplitudes, and uncertainties. Statistical parameters such as mean, variance, skewness, and kurtosis were always applied jointly [66; 64], whereas features relevant to amplitude and uncertainty were sometimes used independently [63; 52]. Moreover, from our review, the energy was the most widely used feature to capture changes of amplitude in EEG signals, and the most common statistical parameters were mean and variance [16].

Focusing on using scalp EEG signals, many detection methods have been developed based on AI and machine learning to automatically detect epileptic seizures in the EEG records [36; 50; 37; 16; 45]. Many studies have utilized single-domain features to capture ictal patterns in EEG signals. Amplitude-integrated EEG (aEEG) was exploited to identify occurrences of high-amplitude seizures [56]. By using an adaptive thresholding method, the method obtained good detection rate (GDR) of 88.50% and FPR/h of 0.18. Nevertheless, this method also responded to artifacts with high amplitudes and required EEG signal beginning with normal EEGs. An energy computed in frequency domain using filter bank analysis and a radial basis function support vector machine (RBF SVM) were jointly employed to characterize the epileptic seizures. As a result, the energies from seizure samples were higher than that of the normal ones. Moreover, the logarithm of variance of DWT coefficients in each sub-band from specific periods of EEGs in a selected channel was used to determine a seizure epoch by thresholding [67]. According to the best result from of each patient, the method obtained the average performances of 93.24% accuracy, 83.34% sensitivity, and 95.53% specificity. Similarly, the author also conducted an experiment using a smaller data set, including only 12 subjects. The results showed that using those features with SVM outperformed a feature combination of line length, nonlinear energy, variance, power, and maximum value of raw EEG signals with the average accuracy, sensitivity, and specificity of 96.87%, 72.99%, and 98.13%, respectively. Furthermore, the STFT spectrogram was used with a modified stacked sparse denoising autoencoder (mSSDA) to detect an epileptic seizure in individual epochs [68]. It was concluded that this method outperformed the other methods conducted in the experiment and obtained the accuracy of 93.82% and F_1 of 96.05%. Nevertheless, no details of chosen data and validation are reported.

On the other hand, a combination of features in a single domain was proposed to capture ictal patterns in many aspects. Fractal dimension called a box-counting dimension (D_B) and an energy were exploited to observe complexity and amplitude of the EEG signal [57]. The records included in [57] were chosen to have the same bipolar montage, and the subject `chb16` was excluded because of the short seizure duration. Eventually, the authors showed that using relevant vector machine (RVM) with these features computed on harmonic wavelet packet transform (HWPT) coefficients potentially achieved the sensitivity of 97.00% and FPR/h of 0.10. Mean, ratio of variance, standard deviation (sd), skewness, kurtosis, mean frequency, and peak frequency were extracted from DWT coefficients [69]. An extreme learning machine (ELM) was employed to classify EEG epochs into a specific class. Due to its effectiveness and efficiency, this combination could accomplish the accuracy of 94.83%. The work in [38] compared the detection performance of using different transformations and different classifiers via the accuracy. First, multi-channel EEG signals were filtered by multi-scale principal component analysis (MSPCA) to remove artifacts. Then, absolute mean value, average power, sd, ratio of absolute mean values, skewness, and kurtosis computed on decomposed signals by EMD, DWT and wavelet packet decomposition (WPD) were applied to random forest (RF), SVM, artificial neural network (ANN), and k-NN for comparisons. Finally, it was concluded that the methods using DWT and WPD obtained 100% accuracy. However, only $2{,}000$ eight-second EEG epochs, $1{,}000$ samples for each group, were selected.

In addition, multi-domain features were also exploited to obtain information in different domains. The work in [70] employed many classifiers: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), polynomial classifier, logistic regression, k-nearest neighbor (k-NN), decision tree, Parzen classifier, and SVM. Variance, root mean squared value (RMS), skewness, kurtosis, and sample entropy (SampEn) were used as time-domain features, and peak frequency and median frequency computed from PSD were exploited to extract information in frequency domain. Combined with a feature selection call LDA with a backward search, the k-NN outperformed the other classifier with the sensitivity of 84.00% and specificity of 85.00%. However, the authors chose only records that contained seizures activities in this study.

Due to the current interest of deep learning and its ability to implicitly extract latent features, many deep learning models were recently applied to classify seizure epochs [71; 72; 73]. Many studies mainly focused on designs and choices of deep learning architectures suitable for indicating seizures [74; 75; 76; 77; 78; 79]. For example, raw EEG signals were purely used as inputs of ANN [80]. It was reported that this method accomplished 100% accuracy; however, the data were specified to contain simple and complex partial epileptic seizures in the frontal area collected from only female subjects.

A bidirectional long short-term memory model (Bi-LSTM) was applied to maximum, minimum, median, mean, sd, variance, skewness, kurtosis, mean absolute deviation (MAD), and RMS extracted from local mean decomposition (LMD) to detect seizure epochs [77]. This method achieved the accuracy of 92.66%, sensitivity of 93.61%, and specificity of 97.00%. A convolutional neural network (CNN)-based ensemble model including InceptionV3 and ResNet152 was designed to classify EEG epochs using the information of PSD [81]. By the chosen records of 11 cases, the accuracy of 92.60%, sensitivity of 92.30%, and specificity of 97.00% were obtained. On the other hand, some attempts tried to transform EEG segments into a specific representation of deep learning model input. For instance, the 224×224 EEG plot images were constructed from raw EEG segments as inputs to CNN models based on VGG16 for seizure classification [82], and the results on their private data set were promising. From current attention on graph neural network, temporal graph CNNs exploiting connections between each channel were designed to detect seizure epochs and determine the seizure areas [83]. EEG data were rearranged to a $5 \times 4 \times M$ tensor according to the channel locations where M is the number of time samples. As a result, the outcomes of the model provided more insights into spatial distributions of seizures.

From the literature, we found that many deep learning models have developed by exploiting characteristics of epilepsy such as ictal patterns, PSD, and connections of each channel to detect epileptic seizures. These models were mostly based on conventional deep learning models, *e.g.*, CNN, ANN, and LSTM; no novel models for time series analysis such as an *attention network* [84] have been applied. In addition, most of them merely focused on the seizure detection performances, and only a few tried to provide more insight into the characteristics of detected seizures such as affected areas.

In conclusion, Tables 3.1 and 3.2 summarize seizure detection methods performed on the CHB-MIT Scalp EEG database by domains of features. We observed that results from epoch-based seizure detection are not realistically suitable for inferring the seizure onset and offset even though those methods achieved promising results. First, false negatives that unfortunately occur in the middle of true positives lead to wrong inferences of seizure events. In this case, a single seizure is incorrectly predicted as many consecutive seizures due to the false negatives. Second, several false positives appearing isolatedly cause too frequent false alarms. Both prediction outcomes are not clinically realistic because there is no abrupt change in EEG patterns during a seizure event. In addition, most of the previous studies intentionally selected data in their experiments to reduce imbalance in the data [85; 73]. However, inclusion and exclusion criteria were not given with exact details; it is not clear how well their methods perform on the excluded data set. It is evident that determining the seizure onset and offset as the first and last epochs of seizures predicted by existing epoch-based seizure classifiers is clinically inappropriate. We, hence, further focus on methods that can potentially indicate the seizure onset and offset so that the seizure predictions are more clinically precise.

3.2 Applications of seizure onset and offset detection

There have been only a few attempts that aim to develop seizure onset and offset detection. One of the first automated seizure offset detection was designed by in [88]. The researchers proposed both patient specific and non-specific algorithms using multi-channel scalp EEG signals. Long EEG signals of patients in the CHB-MIT Scalp EEG database were analyzed by segmenting the signals into five second epochs and advancing each epoch by one second. Both patient specific and non-specific methods used signal energies of 25 contiguous frequency bands spanning 0 - 25 Hz from each channel independently to observe spatial and spectral properties in the epoch. For the patient non-specific setting, a feature vector was constructed from the signal energy averaged over channels of the frequency bands. For the patient-specific case, each feature was a weighted average of the energy of each frequency band over all channels. The weights were calculated based on the differences between the signal energies in ictal and postictal states. Each feature vector was then fed to SVM to classify the epoch as ictal or postictal. A linear SVM was used in the patient-specific case whereas RBF SVM was exploited in the other case. Once the seizure onset had been recognized by the algorithm from their previous study [86], the end of seizure was declared when five consecutive epochs were recognized as postictal. It was reported that the

Table 3.1: Literature summary of automated epileptic seizure detection using the *CHB-MIT* Scalp EEG database when *single-domain* features were used.

Domain	Features	Method	Performance	Ref.
Time	Raw signal	ANN	Acc = 100%	[80]
	aEEG	Thresholding	GDR = 88.50%, FPR/h = 0.18	[56]
	Line length, NE, variance, average power, max	RBF SVM	Acc = 95.17%, Sen = 66.35%, Spec = 96.91%	[67]
	Absolute mean values, average power, sd, ratio of absolute mean values, skewness, kurtosis	MSPCA + EMD + RF	Acc = 96.90%	[38]
		MSPCA + EMD + SVM	Acc = 97.50%	[38]
		MSPCA + EMD + ANN	Acc = 96.90%	[38]
		MSPCA + EMD + k-NN	Acc = 94.90%	[38]
	D_B [4]	RVM	Sen = 97.00%, FPR/h = 0.24	[57]
	Min, max, median ,mean, sd, variance, skewness, kurtosis, MAD, RMS	LMD + Bi-LSTM	Acc = 92.66%, Sen = 93.61, Spec = 91.85%	[77]
Frequency	Energy	RBF SVM	GDR = 96.00%, FPR/h = 0.08	[86]*
	PSD	CNN-based ensemble	Acc = 92.60%, Sen = 92.30%, Spec = 97.00%	[81]
Time-frequency	Spectrogram	STFT + mSSDA	Acc = 93.82%	[68]
	Mean, ratio of variance, sd, skewness, kurtosis, mean frequency, peak frequency	DWT + ELM	Acc = 94.83%	[69]
	Log of variance	DWT + thresholding	Acc = 93.24%, Sen = 83.34%, Spec = 93.53%	[87]
		DWT + RBF SVM	Acc = 96.87%, Sen = 72.99%, Spec = 98.13%	[67]*
	Absolute mean, average power, sd, ratio of absolute mean, skewness, kurtosis	MSPCA[1] + DWT + RF	Acc = 100%	[38]
		MSPCA + DWT + SVM	Acc = 100%	[38]
		MSPCA + DWT + ANN	Acc = 100%	[38]
		MSPCA + DWT + k-NN	Acc = 100%	[38]
		MSPCA + WPD[2] + RF	Acc = 100%	[38]
		MSPCA + WPD + SVM	Acc = 100%	[38]
		MSPCA + WPD + ANN	Acc = 100%	[38]
		MSPCA + WPD + k-NN	Acc = 100%	[38]
	Energy	HWPT[3] + RVM	Sen = 97.00%, FPR/h = 0.25	[57]
	Energy, D_B	HWPT + RVM	Sen = 97.00%, FPR/h = 0.10	[57]

Acc = accuracy, Sen = sensitivity, Spec = specificity, FPR/h = false positive rate per hour
[1] Multi-scale principal component analysis, [2] wavelet packet decomposition, [3] harmonic wavelet packet transform, [4] box-counting dimension
* Use all data records

patient non-specific method was able to detect all seizure ends with an average accuracy of 84% and an average absolute offset latency of 8.9 ± 2.3 seconds while the patient-specific algorithm detected 132 out of 133 seizure offsets with an accuracy of 90.00% and an averaged absolute latency of 10.3 ± 5.5 seconds over patients. However, seizures that slowly changed from the ictal to the postictal periods led to a large delay of seizure offset detection. In contrast, seizure ends were so early detected when the seizure activities were corrupted by artifacts. Additionally, this method requires an onset detection system to alarm the seizure onset first.

Orosco et al. [85] applied stationary wavelet transform (SWT)-based feature extraction in detecting seizures and their onset and offset. Eighteen subjects from the CHB-MIT Scalp EEG database were used to perform patient-specific and patient non-specific scenarios. Non-overlapping two second epochs were decomposed by SWT in each channel individually and coefficients of 4 sub-bands corresponding to normal EEG rhythms were used to extract features. For each channel, mean frequency and peak frequency were calculated from PSD of all selected sub-bands coefficients and a relative energy of each frequency band, an energy of each band normalized by the total energy, was extracted. The features were then spatially averaged over left anterior, right anterior, left posterior, right posterior, and central areas. By feature selection based on the statistical parameter called Lambda of Wilks, 26 features left were applied to LDA and artificial neural network (ANN). The results showed that, in the patient-specific case, LDA outperformed ANN with overall specificity of 99.99%, GDR of 92.60%, false positive rate per hour (FPR/h) of 0.30, and onset and offset latencies of 0.2 and 4 seconds after and before the annotation.

Table 3.2: Literature summary of automated epileptic seizure detection using the *CHB-MIT* Scalp EEG database when *multi-domain* features were used.

Time	Frequency	Time-frequency	Method	Performance	Ref.
Variance, RMS, skewness, kurtosis, SampEn	Peak frequency, median frequency		LDA	Sen = 70.00%, Spec = 83.00%	[70]
			QDA	Sen = 65.00%, Spec = 92.00%	[70]
			Polynomial classifier	Sen = 70.00%, Spec = 83.00%	[70]
			Logistic regression	Sen = 79.00%, Spec = 86.00%	[70]
			k-NN	Sen = 84.00%, Spec = 85.00%	[70]
			Decision tree	Sen = 78.00%, Spec = 80.00%	[70]
			Parzen classifier	Sen = 61.00%, Spec = 86.00%	[70]
			SVM	Sen = 79.00%, Spec = 86.00%	[70]

For the patient non-specific case, LDA also achieved 99.90% specificity, 87.50% GDR, 0.9 FPR/h, and onset and offset latencies of 1.3 and 3.7 seconds, respectively, on average. In this paper, the positive latency was observed when the algorithm detected a seizure before an annotation. Nevertheless, ranges of seizure onset and offset were very wide in both patient specific and non-specific cases. Ranges of the seizure onset and offset in the patient-specific case were 42.4 and 84.4 seconds, while the ranges of the onset and offset in the other case were 248 and 81.3 seconds, respectively. Due to high FPRs per hour obtained from each subject, it is possible only some isolated EEG epochs of each seizure were detected so that GDR was that high.

Another approach focusing on the patient-specific detection of seizure onset and offset that used the CHB-MIT Scalp EEG database was found in [89]. EEG records from 18 patients were analyzed from a one-second sliding window by exploiting an orthonormal triadic wavelet transform. Each EEG epoch was decomposed into specific frequency ranges using triadic wavelets. Statistics-based features were extracted each channel individually from selected frequency bands corresponding to normal EEG rhythms. Then the features of each channel were classified by LDA and k-nearest neighbor (k-NN) independently. Segments which were recognized as seizure for at least 6 channels were marked as 1 representing seizure EEG epochs. The results from the channel-based detection were post-processed by centered moving average (CMA) of length 15 to reduce a false alarm. Eventually, the output from CMA of each epoch was compared to a threshold of 0.4 to determine the final decision. The first epoch detected as seizure was determined as a seizure onset and a seizure end was observed when the final decision changed from 1 to 0, representing transition from a seizure stage to a normal stage. As a result, the method using k-NN achieved 99.62% accuracy, 98.36% GDR, 99.62% specificity, 0.80 FPR/h, 6.32-second seizure onset latency, and -1.17-second seizure offset latency on average. On the other hand, averaged classification performance measurements evaluated by LDA were 98% accuracy, 100% GDR, 98.05% specificity, 4.02 FPR/h, and 1.41 and 8.19 second onset and offset latencies, respectively. This study denoted the positive latency as a time delay that a predicted time point was after an actual time point. However, a seizure offset of some patients was announced 20 seconds after the annotation, whereas a seizure end of other patients was detected 20 seconds prematurely. Furthermore, 100% GDR was accomplished when FPR/h was extremely high. Specifically, FPR/h of some subjects was higher than 10, meaning that there were repeated false alarms about every six minutes.

From the literature, we can conclude that these studies applied conventional machine learning models requiring hand-crafted features to indicate epileptic seizures in EEGs, so the prior knowledge of normal EEGs and epilepsy is needed to extract appropriate features. Next, these studies selected some EEG data from the database, but the inclusion and exclusion criteria were not exactly given. Moreover, according to the reported performance metrics of seizure detection, there is no evidence that the proposed method could determine the duration of seizures accurately, so it is possible that the seizure episodes were correctly detected by the proposed method with only short periods. In addition, mean latencies used in [85; 89] may be misleading for interpreting the seizure onset and offset detection results. In [85], positive and negative latencies were defined as early and late predictions of seizure onset/offset,

respectively, and the mean latency was used as a performance index of delay detection. Cancellations in the calculation of the mean latencies possibly occur, making the average values low even though their ranges could be particularly high.

Chapter IV

PROPOSED METHOD

This chapter presents proposed methods for detection of the starting and ending points of seizure activities. First, we propose a method to indicate significance of individual feature on epileptic seizure detection using the Bayesian error rate in Section 4.1. These significant features are selected as inputs of epoch-based seizure detectors for comparisons with our proposed model. Next, Section 4.2 presents a CNN model to detect seizure episodes using raw EEG segments without any specific knowledge of ictal patterns. Two seizure onset and offset determination methods called a counting-based method and ScoreNet are then explained in Section 4.3. We also establish a new loss function called *log-dice loss* to address a problem of imbalanced-class data. Finally, a metric called an *effective latency index (EL-index)* is introduced to measure the precision of seizure onset and offset detection. Note that we have already published the studies of the work of the feature significance in [16], and the proposed CNN and the counting-based method in [17], so some details here are taken from those articles.

4.1 Feature significance based on Bayesian error rate

A feature is an attribute that measures some properties or characteristics of data. Selecting useful features of a certain problem is important to a classification task. Many studies used only single feature to capture characteristics of seizures in EEGs [90; 63; 42; 47; 52; 48]. However, common feature selection and dimensionality reduction methods used in automatic epileptic seizure detection do not provide information about how a single feature independently improves the classification performances [39; 91; 92]. In this section, we exploit the Bayesian error rate to determine the significance of each feature on the seizure detection. These significant features are further used in epoch-based seizure detection methods to compare classification results with those of the proposed CNN model.

We define x as a feature, $C(x)$ as a class to which feature x is classified, and C_i denotes an actual class i labeled from the data. An error of incorrect classification is generally defined as

$$\text{err} = \int \sum_{C_i \neq C(x)} P\left(C_i|x\right) p(x) dx, \tag{4.1}$$

where $P\left(C_i|x\right)$ is the posterior probability of x in class i, and $p(x)$ is the probability density function of x. Intuitively, the classification error is the total joint probability that the feature is incorrectly classified, and the error depends on both the ability of a classifier and the feature. To find the significance of each feature, we apply the Bayes optimal classifier to indicate the highest contribution of which the feature can possibly achieve. The Bayes optimal classifier provides the minimum error, also known as the *Bayes error rate*, by choosing the class of which the posterior probability is the highest [93]. As a result, the Bayes error rate (err_b) is obtained from

$$\text{err}_\text{b} = \int \sum_{C_i \neq C_\text{max}} P\left(C_i|x\right) p(x) dx, \tag{4.2}$$

where C_max is the class of which the posterior probability is maximum. According to the Bayes' rule, the posterior probability in (4.2) is formulated as

$$P\left(C_i|x\right) = \frac{p(x|C_i)P(C_i)}{p(x)}, \tag{4.3}$$

where $p(x|C_i)$ is a likelihood function, $P(C_i)$ is the prior distribution, and $p(x)$ is the evidence.

In practice, the distribution of the likelihood function is naturally unknown. In order to approximate the distribution, a non-parametric distribution estimation is exploited. For all real x values, the likelihood function can be estimated using the non-parametric kernel smooth function by

$$p\left(x|C_i\right) = \frac{1}{N_i h} \sum_{j=1}^{N_i} K\left(\frac{x - x_j}{h}\right),$$

where N_i is the sample size of class C_i, x_j is a sample in the class, K is the kernel function and h is a bandwidth [94; 95]. Regarding the classification problem, the prior $P\left(C_i\right)$ is assumed to be binomial estimated by the size of class C_i divided by the total number of samples. Finally, $p(x)$ is obtained by the total probability, and it completes the calculation of $P\left(C_i|x\right)$ in (4.3).

Our problem is a two-class classification (normal/seizure) problem, so the expression of the Bayesian error reduces to

$$\text{err}_\text{b} = \int \min_{i=1,2} P\left(C_i|x\right) p(x) dx, \tag{4.4}$$

where C_1 and C_2 stand for the abnormal and normal classes. To evaluate the significant of individual features, we propose to use an improvement rate (**rate**) from a standard condition (err_0) as follows:

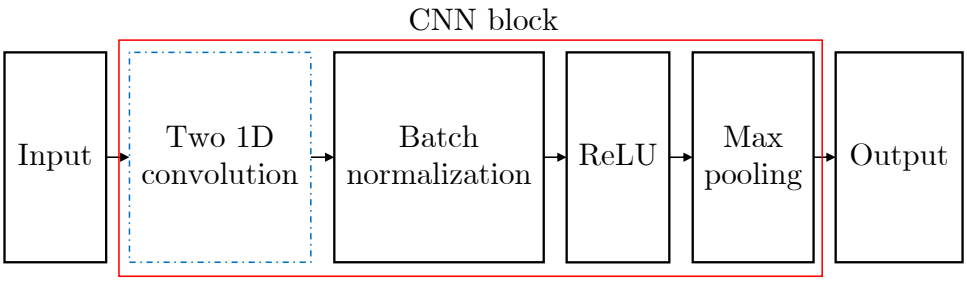
$$\text{err}_0 = \int_{-\infty}^{\infty} P\left(C_2|x\right) p(x) dx = P\left(C_2\right), \quad \textbf{rate} = \frac{\text{err}_0 - \text{err}_\text{b}}{\text{err}_0} \times 100\%. \tag{4.5}$$

The standard condition can be interpreted as classifying all samples to the normal class. This means that the improvement rate on the Bayesian error rate indicates the best contribution of individual feature to the epileptic seizure detection. The improvement rate will be used to analyze features in Section 6.1 and select features for other seizure detectors to compare seizure detection performances with our proposed method in Section 6.2.

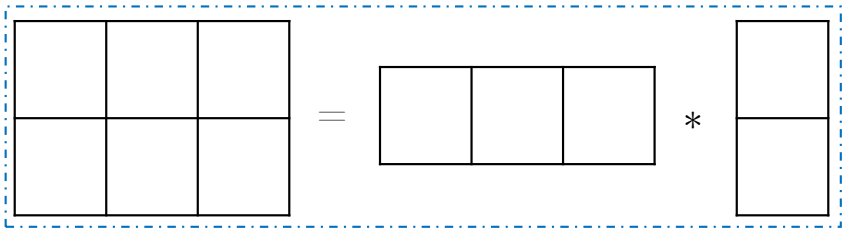
4.2 Convolutional neural network-based seizure detection

In this research, we design a deep CNN model to implicitly extract features from temporal and spatial domains in EEG epochs. The deep model contains CNN blocks for feature extraction and fully-connected layers for classification. The CNN block consists of layers including a convolutional layer, a normalization layer, an activation layer, and a max-pooling layer as shown in Figure 4.1a. Every block has the same sequence of layers but hyperparameters of some layers are changed to serve a physical meaning. For example, some block consists of a one-dimensional max-pooling layer to down sample feature maps in the temporal domain only, whereas a two-dimensional max-pooling layer is used to reduce the dimensions temporally and spatially.

In the CNN block, the convolutional layer is established to capture both temporal information, EEG pattern, and spatial characteristics and dispersion of electric field. To appropriately design the layer, the dimension of EEG epoch is taken into consideration. Suppose that a raw EEG epoch is expressed as a matrix of size $m \times N$, where m is a number of channels, and N is a number of temporal samples in the epoch. Practically, the epoch width is extremely *larger* than its height ($m \ll N$), and ictal patterns are primarily observed by temporal characteristics of EEGs. Therefore, the width of the filter is supposed to be *larger* than its height to capture the temporal patterns rather than the spatial ones. Moreover, we exploit the concept of filter decomposition to reduce a model complexity and to overcome overfitting problems [96]. A two-dimensional filter is decomposed into two one-dimensional filters as shown in Figure 4.1b. The first filter can be physically interpreted as a feature extractor in temporal domain, and



(a) Block of CNN containing convolutional layer, batch normalization layer, activation layer, and max-pooling layer.



(b) Example of filter factorization from a three-by-two filter into three-by-one and one-by-two filters. The first filter aims to extract a temporal feature and the second filter indicates a spatial relationship.

Figure 4.1: Design of CNN block. In the blue box, the two-dimensional filter is factorized into two one-dimensional filters.

the other filter is to find a relationship of a feature between channels. Next, a batch normalization layer is added to reduce an internal covariate shift [29] in the extracted feature. Following the normalization layer, the ReLU function is used as an activation function to fasten the learning procedure [31]. Subsequently, a max-pooling layer is used to draw the most active values of features in a receptive field. The number of blocks is set to appropriately extract high-level useful features from EEG patterns. Finally, dropout layers are applied to reduce overfitting problems, and fully-connected layers are exploited in the last layers to classify each EEG epoch into a specific class (normal/abnormal).

According to this concept, we propose a CNN model illustrated in Figure 4.2 where the model input is a raw EEG epoch, and the model output is a seizure probability. From the shape of the input, we set the first and second filter sizes to 1×3 and 2×1, respectively, to implicitly extract temporal information rather than to observe a spatial characteristic of the adjacent channel. In Figure 4.2, each rectangular box represents a layer, and the description in the box explains the type of the layer. In this case, Conv(h, w, f) is a convolutional layer containing f h-by-w filters, BN stands for a batch normalization layer, ReLU is an activation layer using the ReLU function as the activation function, Max(h, w) is an h-by-w max-pooling layer, Dropout(α) is a dropout layer with the disconnection fraction of α to the input nodes, and FC(n) is a fully-connected layer with n neurons. Totally, there are 17,297 parameters in this model to be trained. The optimizer called ADADELTA is exploited to train the model because it is robust to noise and has an adaptive learning rate [97]. Furthermore, the binary cross-entropy is used as a loss function, and the sample is identified as ictal when the seizure probability is higher than 0.5.

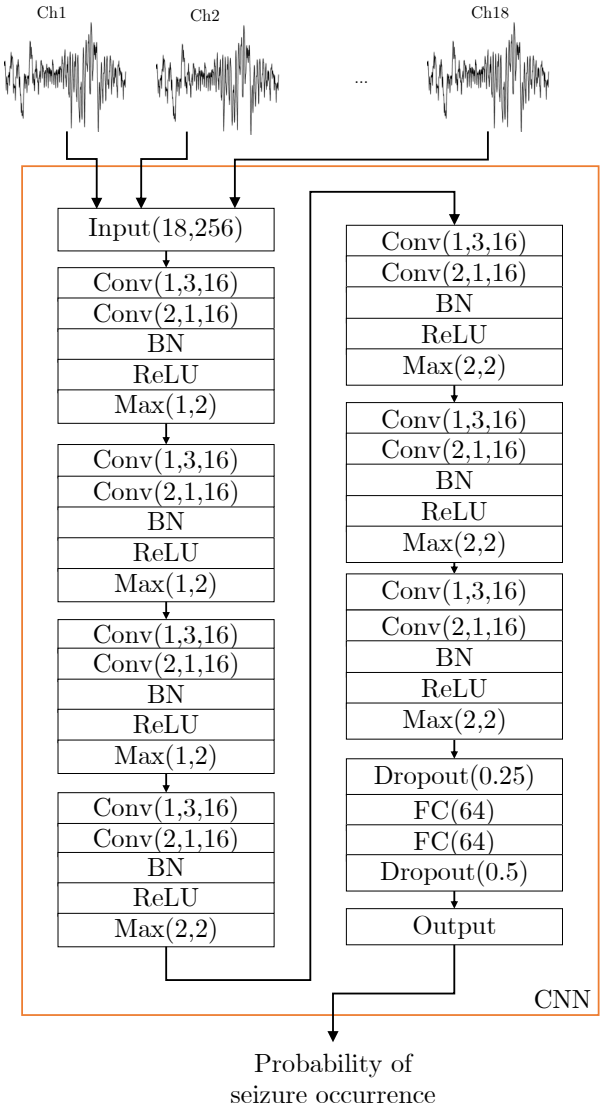


Figure 4.2: Deep CNN structure proposed in this dissertation. Raw EEG segments of 18 channels are inputs, and seizure probabilities are outputs.

4.3 Seizure onset and offset detection

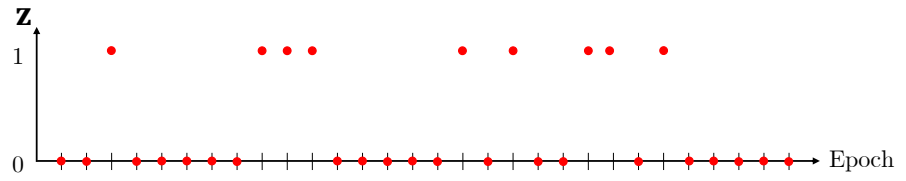
The process of detecting seizure onset and offset from detected seizure episodes is important since predictions of epoch-based seizure detection cannot be directly used to infer the seizure onset and offset. We design two seizure onset and offset detection models, namely, a counting-based method [17] and a neural network-based model called *ScoreNet*. The counting-based method is a criterion-based method initiated to verify our hypothesis that there are isolated false positives making high false positive rate per hour and isolated false negatives leading to wrong inferences of seizure episodes. This method is then formalized to a general form of convolution, and ScoreNet is established by adding flexibility to the convolution form. Additionally, a log-dice loss is proposed as an objective function for training ScoreNet to deal with the imbalanced data problem. We will show in Section 6.2 that using only an epoch-based seizure detector commonly produces numerous false positives and false negatives, and our proposed

detectors of seizure onset and offset can considerably reduce these false predictions.

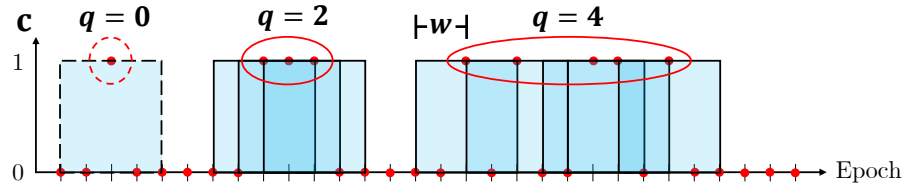
Counting-based model

The counting-based method is a criterion-based seizure onset and offset detector designed from clinical characteristics of seizures. In practice, a seizure does not occur only a few seconds and suddenly disappears [4]. This means that classifying EEG epochs independently is not appropriate for indicating the seizure onset and offset because consecutive epochs are correlated. Moreover, in clinic, it is typical to label some adjacent seizure episodes and normal EEG epochs during those episodes as one seizure. From these characteristics, we establish this method to both merge near detected epochs into one event and disregard isolated predictions that are likely to be false positives.

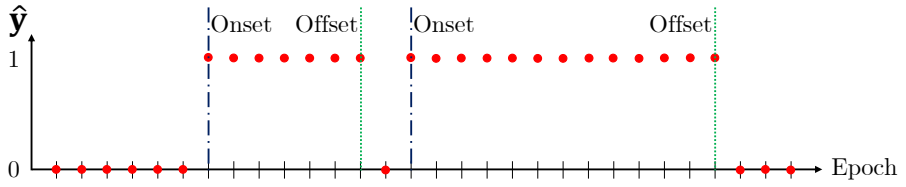
Recall that $z = (z_1, z_2, \ldots, z_N)$ and $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$ are sequences of epoch-based seizure classification outputs and of onset and offset detector outputs of length N, respectively. In this method, z_i can be only either 0 (normal) or 1 (seizure), as shown in Figure 4.3a. We denote k_i as an index of an epoch classified as a seizure, *i.e.*, $z_{k_i} = 1$, for $i = 1, 2, \ldots, m$ when m is the number of seizure-detected epoch, and $1 \leq k_i \leq N$. From the seizure characteristics described above, we can imply that if there are adequately near detected episodes, then all epochs, including the normal ones, among these episodes should be grouped together and identified as ictal. On the other hand, those episodes are regarded as normal when the number of the consecutive detected epochs is insufficient. These conditions can be mathematically written as follows.



(a) Epoch-based classification outputs.



(b) Seizure candidate groups represented by rectangular windows.



(c) Seizure onset and offset detection outcomes. The first candidate group is neglected because $q < p$.

Figure 4.3: Illustration of the counting-based method where $w = 2$ and $p = 2$.

Suppose that q_i is the number of neighboring detected epochs. From the statements above, q_i is counted when k_{j+1} and k_j should be close enough, *i.e.*, $|k_{j+1} - k_j| \leq 2w$ when $2w$ is the maximum

required distance between two adjacent epochs. Therefore, q_i can be mathematically expressed as

$$q_i = \begin{cases} \sum_{j=1}^{i+p-1} \Theta\left(2w - |k_{j+1} - k_j|\right), & i \leq p \\ \sum_{j=i-p}^{m-1} \Theta\left(2w - |k_{j+1} - k_j|\right), & i \geq m - p \\ \sum_{j=i-p}^{i+p-1} \Theta\left(2w - |k_{j+1} - k_j|\right), & \text{otherwise,} \end{cases} \tag{4.6}$$

where $\Theta(x)$ is the Heaviside step function, and p is a required number of neighboring detected epochs. For $-w \leq l \leq w$ and $1 \leq k_i + l \leq N$, the output of the method is

$$\hat{y}_{k_i+l} = \begin{cases} 1 & q_i \geq p \\ 0 & \text{otherwise.} \end{cases} \tag{4.7}$$

Finally, the seizure onset and offset are indicated from the first and last epochs of that group, respectively.

The process of this method illustrated in Figure 4.3 can be firstly interpreted as covering all detected epochs by a window of size $2w + 1$ centered at the epoch; we call the covered epochs *seizure candidates* c since they are potential to be part of a seizure event. Subsequently, groups of consecutive seizure candidates containing at least p pairs of near detected epochs ($z = 1$) are finally declared as ictal, and those not satisfying the condition in (4.7) are disregarded as normal. As shown in Figure 4.3, there are some isolated detected epochs occurring from the epoch-based classification outputs, which is impractical. After applying the counting-based method, the first candidate group is regarded as normal because there is only one detected epoch occurring. On the other hand, the other groups are classified as ictal since there are enough detected epochs within these groups, and the onset and offset of each detected event are the first and last epochs of the group as shown in Figure 4.3c.

Formalization of counting-based method

We notice that the counting-based method is hardly analyzed in the form given in (4.7). Both c_i and q_i are calculated by naively finding near detected epochs. Moreover, the optimal parameters w and p can be obtained by only manual tuning. Hence, we formalize each statement of the counting-based method in a convolution form so that this method can be analyzed easily.

The process formalized in the form of convolution is illustrated in Figure 4.4. Recall that a seizure candidate c_i, an epoch that is near any detected epoch can be alternatively expressed by applying a filter of one of length $2w + 1$ followed by setting one as a threshold:

$$c_i = \Theta\left(\mathbf{z}_i^T \mathbf{1} - 1\right), \quad i = 1, 2, \ldots, N, \tag{4.8}$$

where $\mathbf{z}_i = (z_{i+w}, \ldots, z_i, \ldots, z_{i-w})$, and $z_j = 0$ for $j \leq 0$ and $j > N$, and $\mathbf{1}$ indicates the vector of ones with a compatible size. A group of seizure candidates is then formed from adjacent seizure candidates that have the same value as shown in Figure 4.4. Unlike the counting-based method described above that $c_i = 1$ is only used for grouping, $c_i = 0$ and $c_i = 1$ are both utilized to form groups of seizure candidates.

Subsequently, a *score* s_i is defined as a value that reveals an existence of a pair of positive predictions. Similar to (4.8), s_i can be mathematically expressed as

$$s_i = \Theta\left(\mathbf{z}_i^T \mathbf{1} - 2\right), \quad i = 1, 2, \ldots, N.$$

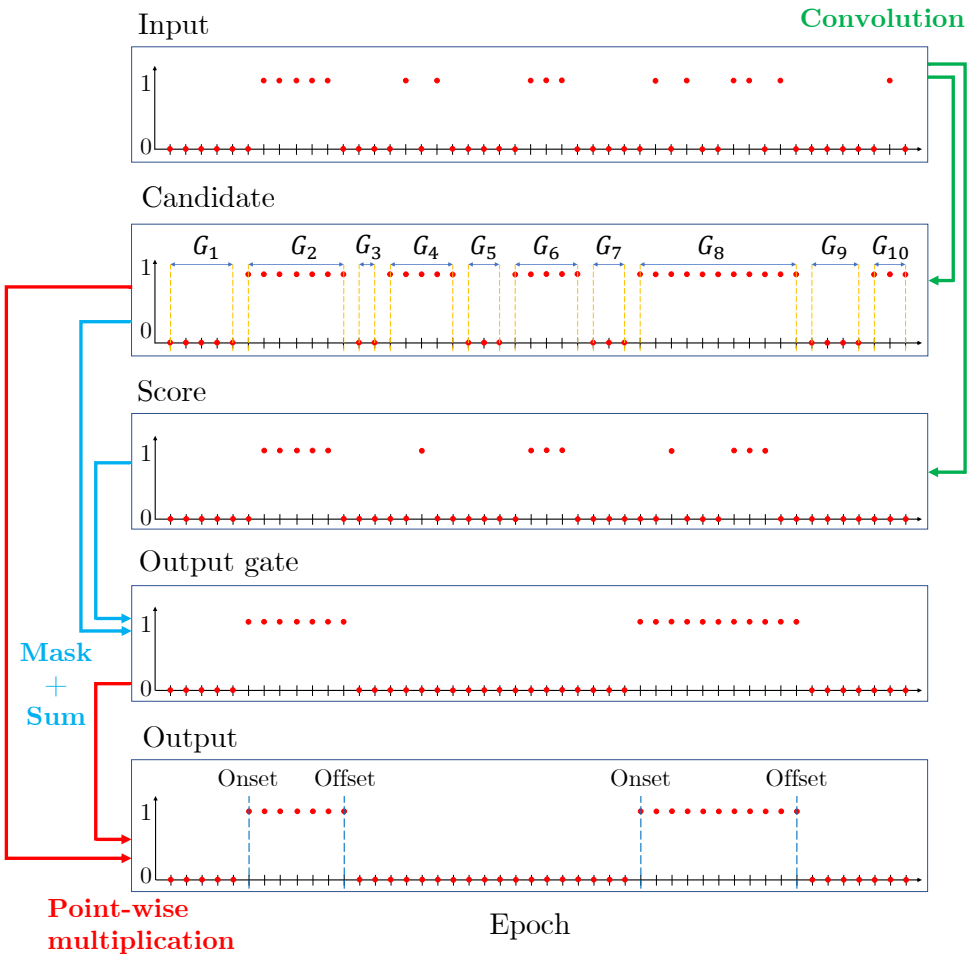


Figure 4.4: General form of the counting-based method when $w = 1$ and $p = 4$. Only groups G_2 and G_8 having scores higher than p are finally predicted events.

The bias term refers to an existence of the prediction pair. We then define an *output gate* o_i which automatically allows a seizure candidate to be an epoch of a seizure when the total score of the group is higher than p:

$$o_i = \Theta\left(\sum_{j \in G_l} s_j - p\right), \quad \forall i \in G_l.$$

when G_l denotes the set of indices of the group l, and p stands for the required number of adjacent prediction pairs. Note that the total score here is similar to q_i in (4.6). We can obviously see that, for any $i, k \in G_l, o_i = o_k$ because the scores are from the same candidate group. Hence, we denote an output gate of the group l as

$$\tilde{o}_l = \Theta\left(\sum_{j \in G_l} s_j - p\right) \quad l = 1, 2, \ldots, m, \tag{4.9}$$

when m is a number of candidate group. Intuitively, $\tilde{o}_l$ is similarly interpreted as an output gate in LSTM; $\tilde{o}_l$ automatically indicates whether a candidate group l is chosen. Finally, for $i \in G_l$ and $l = 1, \ldots, m$,

the output $\hat{y}_i$ is

$$\hat{y}_i = c_i \tilde{o}_l. \tag{4.10}$$

This means that the output $\hat{y}_i = 1$ when the epoch has at least one adjacent detected epoch ($c_i = 1$), and the number of prediction pairs meets the requirement ($\tilde{o}_l = 1$).

ScoreNet

ScoreNet is a neural network-based post-processing model used for inferring the seizure onset and offset. According to the details of the formalized form, it can be interpreted as the convolution of $\mathbf{z}_i$ and a constant filter $\mathbf{1}$. This means that each epoch-based seizure result has an equal effect on the final decision. We hypothesize that results of different epochs may contribute significance differently. Hence, ScoreNet is proposed by adapting the formalized form of the counting-based method.

In ScoreNet, the term $\mathbf{z}_i^T \mathbf{1} - 1$ is generalized as $\mathbf{z}_i^T a + b$ to add more flexibility to this scheme when a is a real-valued vector and b is a real number. With this expression, a and b are allowed to vary upon data. Furthermore, the Heaviside step function must be changed to a differentiable function. In this case, a sigmoid function (σ) is used to determine the candidate c_i. On the other hand, the hyperbolic tangent function (tanh) is applied to calculate the score s_i because its larger output range helps distinguish normal and abnormal groups of seizure candidates. Therefore, we propose the formulae of the candidate and score as

$$c_i = \sigma\left(\mathbf{z}_i^T a_1 + b_1\right), \quad i = 1, 2, \ldots, N, \tag{4.11}$$

and

$$s_i = \tanh\left(\mathbf{z}_i^T a_2 + b_2\right), \quad i = 1, 2, \ldots, N, \tag{4.12}$$

where a_1, a_2, b_1, and b_2 are model parameters. Here, z_i can be a seizure probability, *i.e.*, $z_i \in [0, 1]$, or a binary prediction, *i.e.*, $z_i \in \{0, 1\}$ from the epoch-based classification step. Note that the expressions of c_i and s_i resemble node equations in neural networks that take a linear combination of the inputs and pass through a nonlinear activation. Candidates are then separated into groups using a threshold γ. Since the sigmoid function is a one-to-one function, the bias b_1 is uniquely adapted for a chosen γ during training. Hence, γ can be arbitrarily set to any real value in $(0, 1)$; here, we set $\gamma = 0.5$.

We use an average of the group scores in the calculation of the output gate $\tilde{o}_i$ of the group l in (4.9) instead of the summation to eliminate the effect of the candidate group size. From (4.9), adding parameters a_3 and b_3 as scaling and bias terms, the output gate of the group l is proposed as

$$\tilde{o}_l = \sigma\left(\frac{a_3}{N_l} \sum_{j \in G_l} s_j + b_3\right), \tag{4.13}$$

where N_l is the group size. Finally, the output form in (4.10) is also modified because $\hat{y}_i$ could be low even though c_i and $\tilde{o}_l$ are both sufficiently high. For instance, when $c_i = 0.7$ and $\tilde{o}_l = 0.7$, this epoch should have been classified as abnormal, but it is decided to be normal since $c_i \tilde{o}_l$ is $0.49 (< 0.5)$. Similar to (4.13), we add a weight a_4 and a bias b_4 to scale the value of $c_i \tilde{o}_l$. For any $i \in G_l$, the output $\hat{y}_i$ is given by

$$\hat{y}_i = \sigma\left(a_4 c_i \tilde{o}_l + b_4\right). \tag{4.14}$$

As a result, we have proposed a more general model of seizure onset and offset detection, described by (4.11) - (4.14). Note that the expressions of c_i and s_i resemble node equations in neural networks that take a linear combination of inputs and pass the combination through a nonlinear activation. We propose to train ScoreNet using a nonlinear conjugate gradient method because the method empirically converges faster than other methods when training neural networks [98].

Log-dice loss

We propose the *log-dice loss* for training ScoreNet to overcome an imbalanced data problem. From the soft-dice loss in (2.17), penalties of wrongly classifying hard and easy samples, samples incorrectly classified with large and small errors, are linearly proportional. Costs in incorrectly predicting hard and easy samples are slightly different. We then propose the log-dice loss that includes *different penalties* on the hard and easy samples using the dissimilarity of y_i and $h(x_i)$ in the log-scale.

By conceptually maximizing F_1, from DSC in (2.16), we substitute $\sum_i y_i \log(1-h(x_i))$, $\sum_i (1-y_i)\log(1-h(x_i))$, and $\sum_i y_i \log h(x_i)$ for TP, FP, and FN, respectively. These terms are large when y_i and $h(x_i)$ agree with their original definitions, and small otherwise. For instance, when $y_i = 1$ and $h(x_i) \approx 1$, which is a true positive case, $y_i \log(1-h(x_i))$ approaches infinity, $(1-y_i)\log(1-h(x_i)) = 0$, and $y_i \log h(x_i) \approx 0$. With these definitions, the log-dice loss is proposed as

$$\mathcal{L}_{\text{logDL}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{2\sum_{i=1}^{N} y_i \log(1-h(x_i))}{\sum_{i=1}^{N} [(1+y_i)\log(1-h(x_i)) + y_i \log h(x_i)]} \tag{4.15}$$

The log-dice loss is equal to $1 -$ DSC when substituting TP, FN, and FP in (2.16) for $-\sum_i y_i \log(1-\hat{y}_i)$, $-\sum_i y_i \log \hat{y}_i$, and $-\sum_i (1-y_i)\log(1-\hat{y}_i)$, respectively. Figure 4.5 shows the modified classification indices; by incorporating the logarithmic function, the index value rapidly increases where $\hat{y}$ is close to y.

The value of $\mathcal{L}_{\text{logDL}}$ is in the range of $(0, 1]$ and decreases as y and $\hat{y}$ become more similar. $\mathcal{L}_{\text{logDL}}$ reaches its maximum of one (worst score) under two cases: i) when $y = 0$ (all samples are normal), regardless of the prediction $\hat{y}$ because the index does not consider TN; or ii) when $y = 1$ and $\hat{y} = 0$ (no TP in the prediction). Figure 4.6 compares the cross-entropy, soft-dice, squared-dice and log-dice loss functions as $\hat{y}$ varies under two values of y (one-sample case for illustration). When $y = 0$, $\mathcal{L}_{\text{logDL}}$'s constant loss means that the normal class are neglected when optimizing model parameters. On the other hand, when $y = 1$, and $\hat{y} \leq 0.5$, the log-dice loss has a higher penalty than the cross-entropy, soft-dice and squared-dice losses, implying that $\mathcal{L}_{\text{logDL}}$ optimizes model parameters to prevent FN better than the other losses.

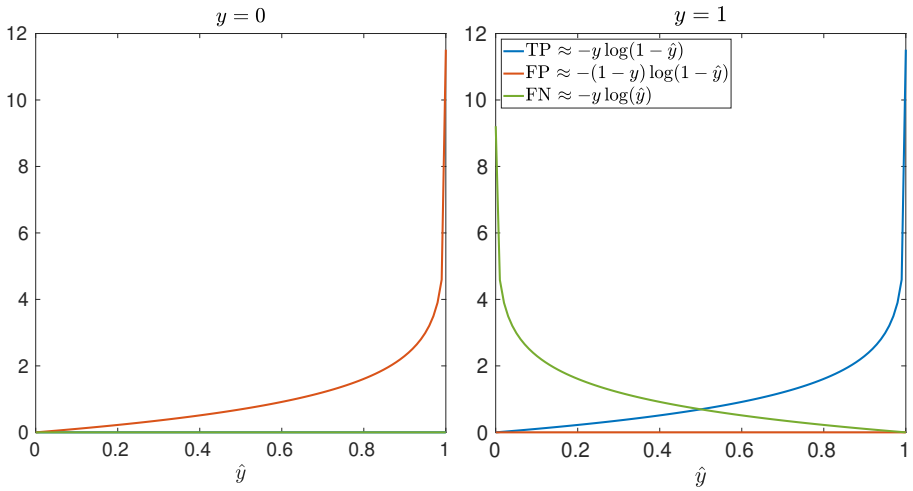


Figure 4.5: Modified classification indices in the log-dice loss.

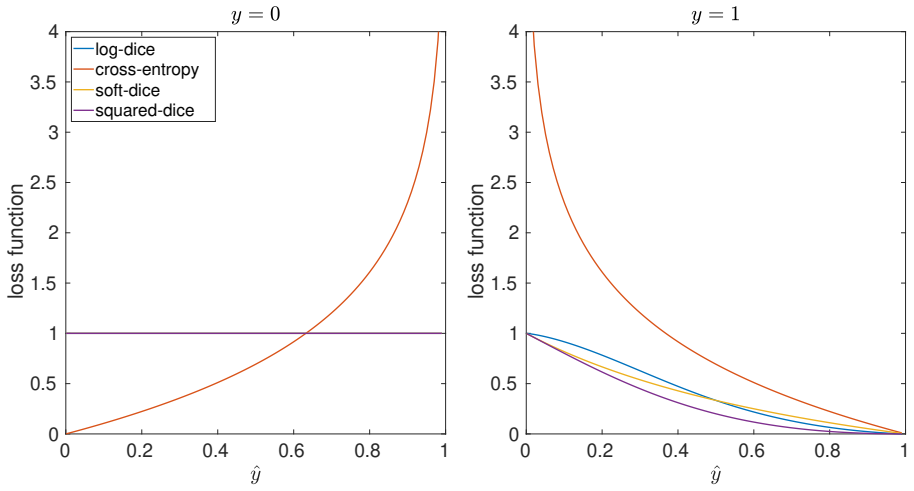


Figure 4.6: Loss functions in binary classification and the proposed log-dice loss.

4.4 Effective latency index

An *effective latency index (EL-index)* is an indicator for measuring delays of correct event detection while undetected events are also taken into account. The motivation is to give a positive score to any event which is correctly detected, and zero for any undetected event. This score should be low when the delay of the detected event is high, and vice versa. Normally, a latency is the time difference between a predicted event (onset/offset) and an actual event. There is currently no consensus of latency determination when situations such as late or early detection and multiple predictions happen. We denote $d_i > 0$ and $d_i < 0$ the latencies of early and late detection of seizure onset and offset, respectively, and there are n seizure activities. Note that d_i is not defined for an undetected seizure event. Moreover, let k_i be an indicator of the event i being detected: $k_i = 1$ when the event is correctly detected, and $k_i = 0$ otherwise. With these notations, the *mean latency* is given by

$$\bar{d} = \frac{\sum_{i=1}^{n} k_i d_i}{\sum_{i=1}^{n} k_i},$$

and the *mean absolute latency* is similarly expressed as

$$\hat{d} = \frac{\sum_{i=1}^{n} k_i |d_i|}{\sum_{i=1}^{n} k_i}.$$

In general, $\bar{d}$ and $\hat{d}$ are particularly low when delays are small, which commonly refer to high precision in the detection. However, both of them ignore all undetected events from the calculation, leading to misunderstanding when their values are promising. For instance, they are low if a few events are almost perfectly detected while the other events are all overlooked. Moreover, $\bar{d}$ is possibly low because of the cancellation of positive and negative latency. Therefore, we propose the EL-index to solve these problems. The EL-index is defined as

$$\text{EL-index} = \frac{1}{n} \sum_{i=1}^{n} k_i r^{|d_i|}, \tag{4.16}$$

where $0 < r < 1$ and n is the number of true seizure events. This decay rate r can be arbitrarily specified by the user according to an impact degree of the delay in each application. For example, in the case of detecting life-threatening abnormalities in the ECG, r should be low because the delay impact is considerable.

The EL-index values range from zero (missing all events) to one (perfectly detecting all events). Given a correct detection ($k_i = 1$), the value $r^{|d_i|}$ exponentially decreases when $|d_i|$ increases, meaning that a higher latency dramatically causes a smaller EL-index. On the other hand, when an event is undetected ($k_i = 0$), the term $r^{|d_i|}$ is ignored from the EL-index. This is similar to assigning $d_i = \infty$ for the undetected event, making $r^{|d_i|}$ approaches zero. Hence, undetected events always provide the least contribution to the EL-index.

If we denote GDR (a good detection rate) a portion of correctly detected seizure events in one record and hence given by $(1/n)\sum_{i=1}^{n} k_i$, we can see that the EL-index given in (4.16) can be regarded as an exponentially weighted GDR. Moreover, an upper limit and a lower limit of the EL-index can be derived as

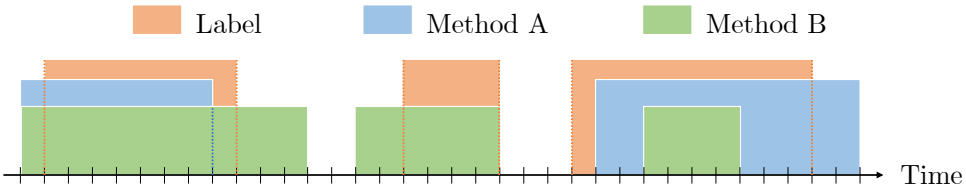
$$\text{GDR} \cdot r^{|d|_{\max}} \leq \text{EL-index} \leq \text{GDR},$$

where $|d|_{\max}$ is the maximum value of absolute time delays. It is evident that the EL-index cannot be higher than GDR, and it is bounded below by the function $\text{GDR} \cdot r^{|d|}$. Moreover, the EL-index can distinguish distributions of collected time delays. Suppose we have two cases of time delay samples from test results. The distributions of these two cases are narrowly spread and highly varied, but the samples of two cases have the same mean absolute latency $\hat{d}$ and the same GDR. When all time delays are similar, *i.e.*, $d_i \approx \pm\hat{d}$, for any detected event i, we can derive that the EL-index is close to the exponential bound

$$\text{EL-index} \approx \text{GDR} \cdot r^{\hat{d}}. \tag{4.17}$$

In contrast, when the time delays are highly varied, the EL-index is far from the exponential bound in (4.17), and the EL-index in this case is always higher than that of the first case:

$$\text{GDR} \cdot r^{\hat{d}} \leq \text{GDR} \cdot \frac{\sum_{i=1}^{n} k_i r^{|d_i|}}{\sum_{i=1}^{n} k_i} = \frac{1}{n}\sum_{i=1}^{n} k_i r^{|d_i|}.$$



		Mean latency	Mean absolute latency	EL-index
Onset	Method A	0.00	1.00	0.60
	Method B	0.00	2.00	0.81
Offset	Method A	0.50	1.50	0.57
	Method B	0.00	2.00	0.82

Figure 4.7: Comparisons of time-based measurements.

Figure 4.7 shows comparisons of these three metrics when $r = 0.9$. It is obvious that the mean latency is not an appropriate metric because the mean latency of onset latencies from both methods are

zero, but both methods do not perfectly detect seizure onsets. Method A obtains less mean absolute latency of onset than method B but method A cannot detect the second seizure event ($k_2 = 0$). The EL-index of detecting seizure onset of method B is, therefore, higher than that of method A. Moreover, the seizure onsets are more accurately detected more than the seizure offsets when method A is considered; the EL-index of detecting the onset is higher. For method B, latencies of detecting the event terminations are relatively diverse compared to those of detecting the starting points, so the EL-index of detecting offset is higher than that of detecting onset given the same mean absolute latency.

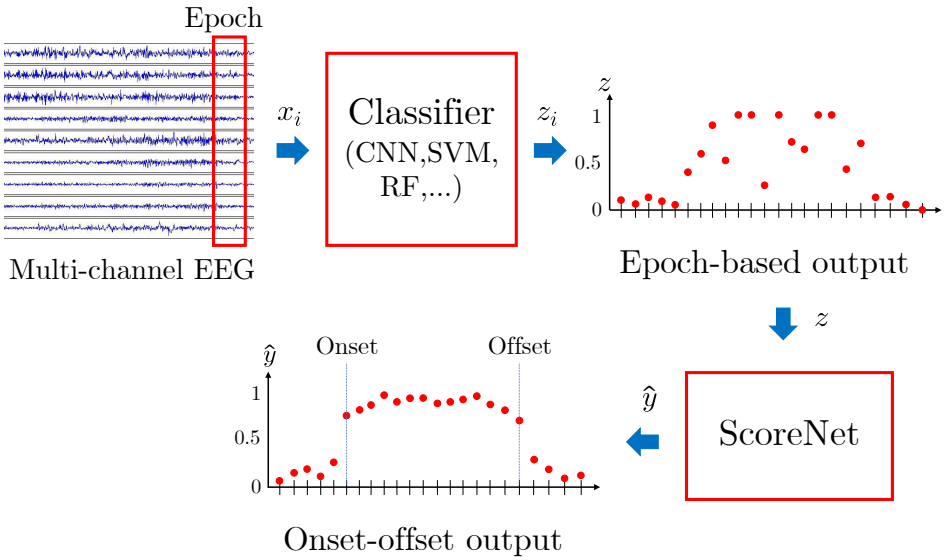


Figure 4.8: Scheme of using ScoreNet with any classifier.

In summary, we propose the CNN model in Section 4.2 to detect seizure episodes in EEG epochs and two seizure onset and offset detector to indicate the onset and offset in Section 4.3. ScoreNet can be used with any epoch-based classifiers as shown in Figure 4.8. Given gradients of ScoreNet in Appendix B, we optimize this loss over $a_1, a_2, a_3, a_4, b_1, b_2, b_3$, and b_4 using the conjugate gradient method. Finally, the EL-index is used to measure time delays of detecting the seizure onset and offset.

Chapter V

MATERIAL AND EVALUATION

This chapter explains a material, performance metrics, and an evaluation scheme used in this dissertation. We firstly describe the database called CHB-MIT Scalp EEG that meets the scopes of our work. Next, the performance metrics, including epoch-based, event-based, and time-based metrics, are presented. Finally, a leave-one-record-out cross validation (LOOCV) is described as a validation scheme.

5.1 CHB-MIT Scalp EEG database

According to the scopes of this work, the research scheme is patient-specific; data in training and testing processes are collected from the same patient. The data are multi-channel scalp EEG signals, and there exist at least two EEG records containing an epileptic seizure in each case to ensure anomaly samples in training data. Moreover, the starting and ending time points of each individual seizure must be annotated. From the above conditions, the CHB-MIT Scalp EEG database [35] is the only open-source database that currently meets all requirements, so we select to conduct experiments on this database. The other widely used databases that do not match our scopes are listed with comments as follows.

- Bonn University database [99]: This database consists of both scalp EEG and intracranial EEG collected from a single channel.
- Temple University Hospital EEG Seizure database [100]: None of patients are included in both training and evaluation sets.

The CHB-MIT Scalp EEG database comprises of EEG recordings of 24 cases collected from 23 subjects at the Children's Hospital Boston [35]. Every signal was recorded at the sampling frequency of 256 Hz with resolution of 16 bit. The international 10-20 system was exploited to locate electrodes on the scalp and both referential and bipolar montages were used. In summary, there are 686 long EEG records which include 129 records containing 198 seizures in this database. The total duration and the numbers of seizure activities from each case are concluded in Table 5.1. All records are publicly and freely downloaded from PhysioNet (https://physionet.org/physiobank/database/chbmit/).

5.2 Performance metrics

Performance indices of the seizure onset and offset detection are crucial for comparisons of the detection algorithms. We categorize the performance metrics into three groups regarding the purposes of the evaluations. As a binary classification problem, epoch-based and event-based metrics are commonly used to evaluate the performance of the seizure detection. In addition, time-based metrics, including an EL-index described in Section 4.4, are considered when we assess methods of the seizure onset and offset detection.

Table 5.1: Summary of the CHB-MIT Scalp EEG database.

Cases	# records	Total duration (sec)	# seizures	Seizure duration (sec)
chb01	42	145,988	7	449
chb02	36	126,959	3	175
chb03	38	136,806	7	409
chb04	42	561,834	4	382
chb05	39	140,410	5	563
chb06	18	240,246	10	163
chb07	19	241,388	3	328
chb08	20	72,023	5	924
chb09	19	244,338	4	280
chb10	25	180,084	7	454
chb11	35	123,257	3	809
chb12	24	85,300	40	1,515
chb13	33	118,800	12	547
chb14	26	93,600	8	177
chb15	40	144,036	20	2,012
chb16	19	68,400	10	94
chb17	21	75,624	3	296
chb18	36	128,285	6	323
chb19	30	107,746	3	239
chb20	29	99,366	8	302
chb21	33	118,189	4	203
chb22	31	111,611	3	207
chb23	9	95,610	7	431
chb24	22	76,640	16	527
sum	**686**	**3,536,540**	**198**	**11,809**

Epoch-based metrics

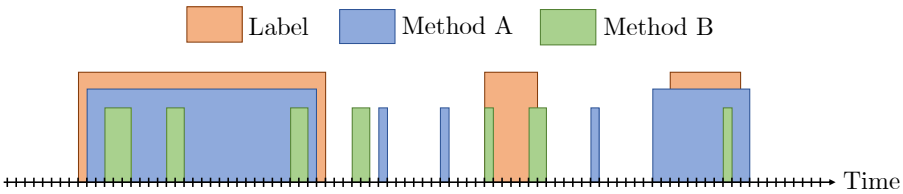
Epoch-based metrics are used to perform an evaluation of the detection performance when each epoch is regarded as a data sample. The calculations of the epoch-based metrics are related to the confusion matrix. For instance, many studies has reported the performance as accuracy, sensitivity, and specificity [36; 42; 101; 16]. The epoch-based metrics can also imply how well the classifier is when a duration is concerned. The epoch-based metrics are calculated from a confusion matrix containing the numbers of true positive (TP), false positive (FP), false negative (FN), and true negative (FN). With these values, many metrics are established for specific purposes. For example, common metrics such as accuracy (Acc), sensitivity (Sen), and specificity (Spec) are defined as

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100\%,$$
$$\text{Sen} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%,$$
$$\text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\%.$$

The accuracy is normally used to indicate the overall performance of the classification, while the sensitivity and specificity are indicators determining the performance of correctly classifying outputs as ictal and normal, respectively. Moreover, F_1, also known as F-measure is the measure of classification performance that takes an imbalance of the data into account [102]. It is calculated from a harmonic mean of precision, or positive predictive value, and recall, or sensitivity. In other words, F_1 can also be calculated as follows:

$$F_1 = \frac{2TP}{2TP + FN + FP} \times 100\%.$$

However, it is hardly said that high values of the epoch-based metrics are clinically referred to promising results. For example, as shown in Figure 5.1, the epoch-based metrics by method A is still incredibly high even though the detector misses one whole short seizure activity when other seizure episodes are correctly classified. Therefore, using only epoch-based metrics is not sufficient to conclude the performance of the detection method.



	Epoch-based metrics				Event-based metrics	
	F_1 (%)	Acc (%)	Sen (%)	Spec (%)	GDR (%)	FPR/h
Method A	80.00	80.65	80.00	81.25	66.67	3
Method B	49.23	64.51	35.56	91.67	100.00	1

Figure 5.1: Comparisons of epoch-based and event-based metrics.

Event-based metrics

Event-based metrics are used to evaluate a classifier based on seizure events in long EEG signals. In this case, the true positive is counted when there is an overlap between predicted seizure episodes and the annotation, the false positive is declared when a detected period of EEG signal does not overlap an actual seizure period, and the false negative is indicated when there is no detected epoch during a seizure activity. Note that there is no definition of true negative for the evaluation.

Two common metrics, namely, good detection rate (GDR) and false positive rate per hour (FPR/h) calculated based on the intersection of detection results and annotations are used in this application [57; 86; 56]. GDR, or event-based sensitivity, is defined as the percentage of detected seizure events that have an overlap with the annotations. FPR/h, also called false detection rate per hour, is the proportion of wrongly identified seizures in one hour. A higher GDR indicates a higher number of correctly detected seizure events, while a small FPR/h refers to having a lower number of wrongly recognized seizure events.

Nevertheless, promising values of event-based metrics should be concerned to avoid being misled into a conclusion of the correct detection when a duration is considered. For instance, as demonstrated by method B in Figure 5.1, detecting an occurrence of seizure at the last second of an actual seizure event is still considered as good detection even though the detection system nearly misses the whole seizure event. Hence, the event-based metrics should be used with the epoch-based metrics for fairness and completeness of the evaluation.

Time-based metrics

Time-based metrics are indicators of measuring time delay of the seizure onset and offset detection. A latency is a measurement of identifying the difference between actual and detected time points. Unfortunately, there is no standard calculation of the latency in this application since many studies have previously defined the latency differently [85; 89]. Therefore, in this study, the latency is defined as a delay of a detected seizure onset/offset when an actual seizure onset/offset is set to be a reference. We use positive and negative latencies defined in Section 4.4; the positive and negative onset/offset latencies refer to the declarations of onset/offset after and before the annotation, respectively.

Normally, the mean latency and the mean absolute latency are used to indicate the precision of detecting the seizure onset and offset [85; 88]. As discussed in Section 4.4, these metrics are inappropriate to measure the performance of seizure onset and offset detection due to the cancellations of positive and negative latencies, and the neglect of undetected seizure events. Therefore, the proposed EL-index is mainly used to quantify the performance, and the mean latency is exploited to determine the trend (early/late) of the detection error.

5.3 Leave-one-record-out cross validation

A means of evaluating the automatic epileptic seizure detection is essential to compare the performances of each model. Since our purpose is to detect seizure episodes and to indicate onsets and offsets, using a validation method that considers EEG epochs as samples is not appropriate because we cannot determine the onsets and offsets if the results are not sequential. In this case, a leave-one-record-out cross validation (LOOCV) is exploited. Suppose that there are n patients, and each subject has k records. For each patient, EEG records are divided into two groups: training set and validation set as shown in Figure 5.2. A training set contains $k-1$ records, and the excluded record is of a validation set. The model is trained on the data in the training set, and then validated on the record in the validation set. Performance metrics described in Section 5.2 are then evaluated from the validation set. This process repeats until every record is in the validation set. Finally, each metric from all $k \times n$ validation cases are accumulated and reported in terms of the mean, median, and interquartile range of these metrics.

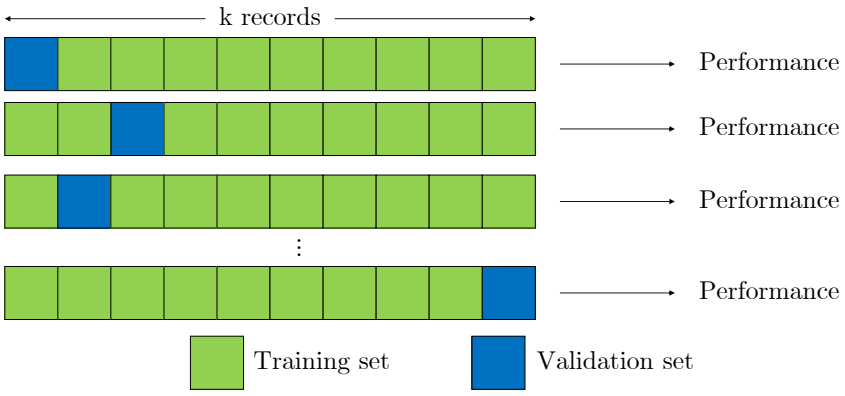


Figure 5.2: Leave-one-record out cross validation scheme.

Chapter VI

EXPERIMENT

This chapter demonstrates primary experiments conducted in this dissertation. Firstly, the significance of each individual feature for discriminating ictal patterns from normal EEGs using the Bayesian error rate is explored in Section 6.1. Next, Section 6.2 provides details of epoch-based seizure classification methods, and reports of seizure onset and offset determination are demonstrated in Section 6.3. Both experimental setups and results of each topic are also included in this chapter, and the results obtained from this work are compared to those reported in other studies with a remark of data used in their experiments. Since some experimental results were already published [16; 17], and some is available in [18], the contents in this chapter are partly taken from these references. In addition, supplementary experiments such as varying parameters of the counting-based method are explained in appendix. Codes of running ScoreNet and experimental results are available at https://github.com/Siyaosk129/ScoreNet.

6.1 Feature analysis

This experiment aims to analyze the significance of each individual feature for the seizure detection using the Bayesian error rate proposed in Section 4.1. Features used in this study are from three domains: time domain, frequency domain, and time-frequency domain. The obtained significant features are then fed to an existing feature selection method called correlation-based feature selection (CFS) [103] to examine redundancy between these features.

Experimental setup

According to Section 4.1, the kernel function K and the bandwidth h need to be specified in the kernel density estimation. In this research, the Gaussian kernel and the optimal bandwidth $h \approx 1.06\hat{\sigma}N_i^{-\frac{1}{5}}$, minimizing mean integrated squared error, were chosen, where $\hat{\sigma}$ and N are the sample standard deviation of the data and the sample size. We randomly selected two records from each case in the CHB-MIT Scalp EEG database subject to the inclusion condition that every record must contain at least one seizure activity. Those random records for assessing the improvement rate of the Bayes error rate of each feature are shown in Table 6.1. Since the EEG channels of each record were originally inconsistent, we firstly modified the EEG channels so that the montages were bipolar. The modified EEG channels used in this study were sequentially listed as follows: *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*, and *P8-O2*.

In this experiment, an EEG epoch was defined by a segment of a *four-second* raw EEG signal in every channel with an overlap of three seconds to the consecutive epoch. These choices were selected from inspecting the processing step using the commercial software called Persyst [104]. A feature was then computed from each channel of the EEG epoch independently. Only widely used features in the literature shown in Table 6.2 are selected for the analysis, and formulae and descriptions of these features are presented in Appendix A.1. A template length m and a tolerance r of which ApEn and SampEn from normal and abnormal EEGs are maximally different were selected. As a result from Appendix C.2, the template length m and the tolerance r were set to $m = 2$ and $r = 0.1$sd, where sd is the sample standard deviation of the EEG data. All time domain features were calculated from a raw EEG signal, whereas features from the frequency domain were extracted from power spectral density, and time-frequency-domain

Table 6.1: List of records used to evaluate feature significance.

Records					
chb01_04	chb01_16	chb02_16+	chb02_19	chb03_03	chb03_35
chb04_08	chb04_28	chb05_06	chb05_13	chb06_01	chb06_04
chb07_13	chb07_19	chb08_02	chb08_05	chb09_06	chb09_08
chb10_38	chb10_89	chb11_92	chb11_99	chb12_33	chb12_38
chb13_19	chb13_55	chb14_04	chb14_18	chb15_06	chb15_15
chb16_17	chb16_18	chb17a_04	chb17b_63	chb18_32	chb18_35
chb19_29	chb19_30	chb20_14	chb20_16	chb21_21	chb21_22
chb22_21	chb22_22	chb23_06	chb23_09	chb24_04	chb24_11

Table 6.2: List of features for the Bayesian error rate evaluation.

Domains	**Features**
Time	Mean, variance, coefficient of variation (CV), skewness, kurtosis, max, min, energy, nonlinear energy (NE), line length, Shannon entropy (ShEn), approximate entropy (ApEn), sample entropy (SampEn), number of zero-crossing, number of local extrema, mobility, complexity
Frequency	Intensity weighted mean frequency (IWMF), intensity weighted bandwidth (IWBW), spectral entropy (SE), peak frequency, peak amplitude
Time-frequency	Mean, absolute mean, variance, skewness, kurtosis, max, min, energy, line length

features were computed from discrete-wavelet-transform (DWT) coefficients with the Daubechies 4-tap wavelet for five levels.

Currently, the EEG channel selection is still an open research problem, and using multi-channel EEG signals may be redundant. Moreover, some commercial software also analyzes the seizure activity over the left side and the right side of the brain instead of using individual channels. For these reasons, we used a spatial averaging of features over the left side and the right side of the brain independently, denoted as x_{left} and x_{right}, respectively. We then estimated the posterior probability distributions and computed the Bayes error rates using the left and right feature representatives.

Results and discussions

From the chosen records, there were $263{,}424$ samples in the normal group and $4{,}677$ epochs belonging to the seizure group, so $\text{err}_0 = 0.0174$. The features with an improvement rate higher than 4.50% were considered as significant. Table 6.3 shows the Bayes error rate and improvement rate of each time-domain feature and frequency-domain feature. The results from the data set showed that most features achieved almost the same Bayes errors which were close to err_0, except for the *variance, energy, nonlinear energy*, and *Shannon entropy* that obtained high improvement rates of 4.77% to 10.07%, where the nonlinear energy reached the highest improvement rate. Figure 6.1 displays the improvement rates

of all time-frequency-domain features in each wavelet decomposition level. Overall, the results showed that the *variance* and *energy* of the wavelet coefficients in all decomposition levels yielded relatively high improvement rates compared to other features. Specifically, *energy* from level D1 of the left half brain accomplished the highest improvement rate of 13.51%. *Line length* from levels D5 and A5, and *kurtosis* from D1 of the left hemisphere also achieved significant improvement rates.

Table 6.3: Bayes error (err_b) and improvement rate of time-domain and frequency-domain features using the CHB-MIT Scalp EEG database.

(a)) **Time-domain** features.

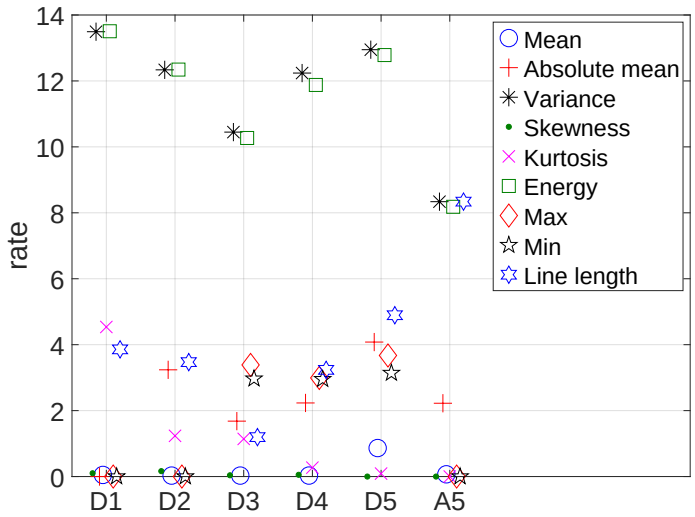
Feature	Left side		Right side	
	err_b	**rate**	err_b	**rate**
Mean	0.0174	0.00	0.0174	0.05
Variance	0.0160	**8.20**	0.0166	**4.78**
CV	0.0174	0.00	0.0174	0.00
Skewness	0.0174	0.00	0.0174	0.00
Kurtosis	0.0174	0.08	0.0174	0.17
Max	0.0174	0.00	0.0174	0.00
Min	0.0174	0.00	0.0174	0.00
Energy	0.0160	**8.18**	0.0166	**4.77**
NE	0.0157	**10.07**	0.0160	**8.36**
Line length	0.0166	1.92	0.0167	1.16
ShEn	0.0174	0.33	0.0165	**5.64**
ApEn	0.0174	0.00	0.0174	0.00
SampEn	0.0174	0.00	0.0174	0.00
Local extrema	0.0174	0.00	0.0174	0.00
Zero-crossing	0.0174	0.09	0.0174	0.09
Mobility	0.0174	0.00	0.0174	0.00
Complexity	0.0174	0.00	0.0174	0.00

(b)) **Frequency-domain** features.

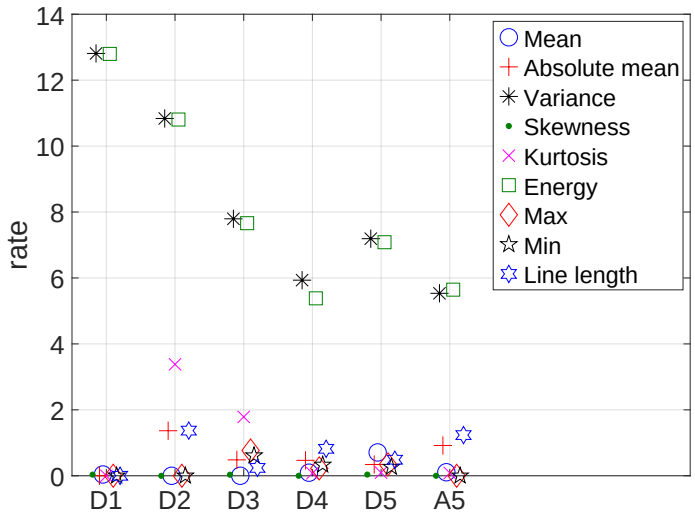
Feature	Left side		Right side	
	err_b	**rate**	err_b	**rate**
IWMF	0.0174	0.00	0.0174	0.00
IWBW	0.0174	0.00	0.0174	0.00
SE	0.0174	0.00	0.0174	0.00
Peak amplitude	0.0174	0.00	0.0174	0.00
Peak frequency	0.0174	0.00	0.0174	0.00

We found that the most significant features were related to amplitudes and variations of the signals, such as variance, energy, and nonlinear energy. Additionally, features that can capture changes in amplitude, frequency, and rhythmicity of EEGs gain some improvement since there is continuous evolution of amplitude, frequency, and rhythms during seizure activities compared to the background [20].

On the other hand, frequency-domain features do not help improve the performance from the baseline because, in this data set, there are artifacts causing the seizure probability to be less than that of the other class.



(a) *left* hemisphere.



(b) *right* hemisphere.

Figure 6.1: Improvement rates based on the Bayesian method of time-frequency domain features. `D1`, `D2`, `D3`, `D4`, `D5`, and `A5` represent sub-bands from which the features are extracted.

Subsequently, the significant features achieving the improvement rate higher than 4.50% (in total, 34 features) were then applied to the CFS algorithm to examine their redundancy; see more details of this algorithm in Appendix A.2. Figure 6.2 shows that the merit score initially increases as m (the number of features in the subset) increases until the feature subset contains five features, with a maximum score of 1.25×10^{-2}. As m increases beyond five, the merit score decreases.

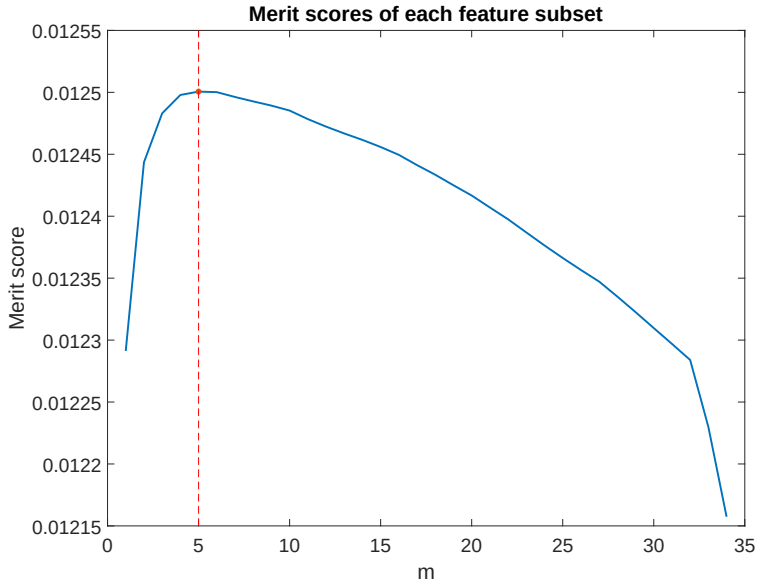


Figure 6.2: Merit scores of feature subsets. The subset size achieving highest merit score is 5.

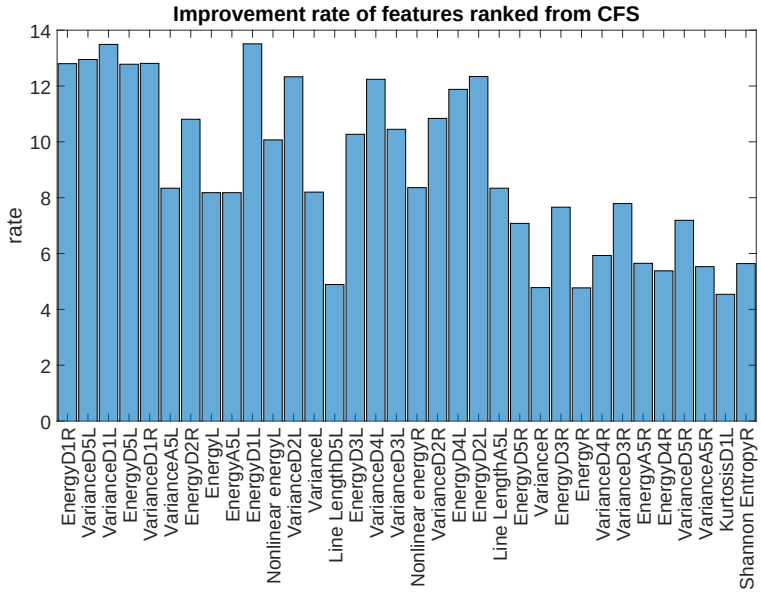


Figure 6.3: The features ranked by CFS and their improvement rates. All features in the optimal subset also obtained high improvement rates.

Additionally, the features and their improvement rates in the final feature subset were ranked in the descending order by CFS as shown in Figure 6.3, where L and R stand for features computed from the left side and right side, respectively. The features in the optimal subset were *EnergyD1R, VarianceD5L, VarianceD1L, EnergyD5L*, and *VarianceD1R*. Accordingly, these five features also achieved relatively high improvement rates among the significant features. As shown in Figure 6.2, the merit scores of the optimal feature subset of each size were not that different. This indicated that the results of feature significance from our experiments also agree with the outputs of the CFS algorithm. Moreover, these

five features obviously require just only $\mathcal{O}(N)$ for the total computation when N is the data size. Hence, these five features should be at least used as features in the automatic epileptic seizure detection because of their high improvement rates and low computational complexity.

6.2 Seizure detection

The goal of this experiment is to indicate an occurrence of an epileptic seizure in an EEG epoch. We design a CNN model to detect the seizure without any feature extraction and noise removal processes. Four classifiers, namely, logistic regression, SVM, decision tree, and random forest are exploited for comparisons epoch-based seizure detection results with those obtained from the proposed CNN model. Moreover, the counting-based method and ScoreNet are applied to these classifiers, including the proposed CNN, to show the consistency of the improvement in the seizure detection performance.

Experimental setup

This experiment was fully performed on *all records* in the CHB-MIT Scalp EEG database in a patient-specific scenario; all EEG data in training and test sets were from the same patient. Similar to Section 6.1, from the inconsistency of EEG montages, the EEG records were firstly rearranged so that the montages were bipolar. The sequential order of the modified 18 channels were *FP1-F7, F7-T7, T7-P7, P7-O1, FP1-F3, F3-T3, T3-P3, P3-O1, FP2-F4, F4-C4, C4-P4, P4-O2, FP2-F8, F8-T8, T8-P8, P8-O2, FZ-CZ*, and *CZ-PZ*. According to LOOCV, the modified EEGs from each case were split into training and validation sets. The long EEG records in each set were segmented into non-overlapped epochs of *one-second* period; the EEG segment was of size 18×256. These epochs were then randomly shuffled to reduce bias and variance from the order of the data.

Table 6.4: List of extracted features used for seizure detection.

Domains	**Features**
Time	Variance, energy, NE, ShEn, ApEn, SampEn
Frequency	Energies from eight sub-bands ranging 0 - 25 Hz
Time-frequency	Mean of absolute value, variance, energy, maximum, minimum, line length

In order to compare the classification outcomes of the proposed CNN model, four classifiers, namely, logistic regression, linear SVM, decision tree, and random forest were applied to indicate seizure episodes from EEG epochs. Note that we attempted to use RBF SVM to determine seizure occurrences. However, it was impossible to select model parameters such that the model can detect at least one seizure from every patient. Thus, RBF SVM was not included in the comparison. Since the chosen classifiers require hand-crafted features to characterize ictal and normal patterns, we chose features that have been widely used and able to discriminate epileptic event from normal epochs [86; 36; 50; 16]. We also exploited the knowledge obtained in Section 6.1 for selecting the features. The final set of features for these classifiers is demonstrated in Table 6.4, where the template length m and the tolerance r were set to $m = 2$ and $r = 0.1$sd where sd is the sample standard deviation of the EEG data. The time-domain features and frequency-domain features were extracted from a raw EEG epoch and its power spectral density, whereas DWT coefficients from five decomposition levels using the Daubechies wavelet of four taps were used to calculate the time-frequency-domain features. Each feature was computed from an EEG signal in each channel, and then normalized to a z-score. Totally, 900 features from every channel

were combined into a feature vector for the classification of EEG epoch. In addition, we used a grid search for tuning hyperparameters of the other epoch-based classifiers. Similar to the CNN case, the hyperparameters that yielded at least one true positive in every patient and achieved high overall F_1 from test records were selected.

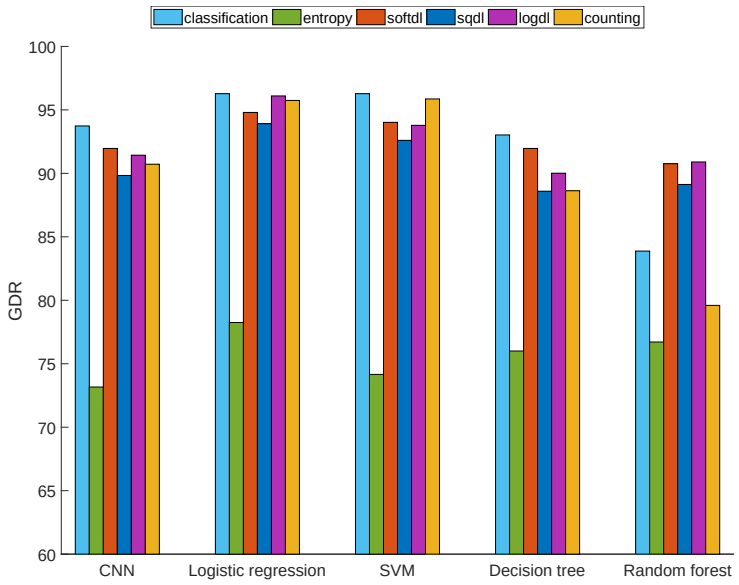
After the epoch-based seizure detection process, the counting-based method and ScoreNet proposed in Section 4.3 were then applied as post-processing techniques to improve the epoch-based classification performance. We firstly collected seizure predictions of each epoch-based classifier from both training and validation sets for learning and testing abilities of the post-processing approaches. These predictions were then arranged by time indices to match the annotations of seizure onsets and offsets. We set parameters of the counting-based method to $w = 6$ and $p = 2$ since these parameters achieved the best F_1 as reported in Appendix C.3. ScoreNet was trained and validated on the arranged predictions of the training and validation set. We also exploited the binary cross-entropy, soft-dice loss [32], and square-dice loss [33] for learning ScoreNet.

Finally, we evaluated and compared the seizure detection methods using epoch-based metrics and event-based metrics to demonstrate seizure detection performance in both aspects. In the epoch-based metrics, F_1 was mainly focused as its suitability of measuring class-imbalanced classification results. The other metrics: accuracy, sensitivity, and specificity, were reported as supplementary outcomes. GDR and FPR/h were considered as the event-based metrics to determine correctness of detecting seizure events. Additionally, we also adopted a patient-specific LOOCV as a validation scheme for the case of ScoreNet. The performance metrics were collected from each validation case and then demonstrated in terms of the mean, median, and interquartile range of these metrics.

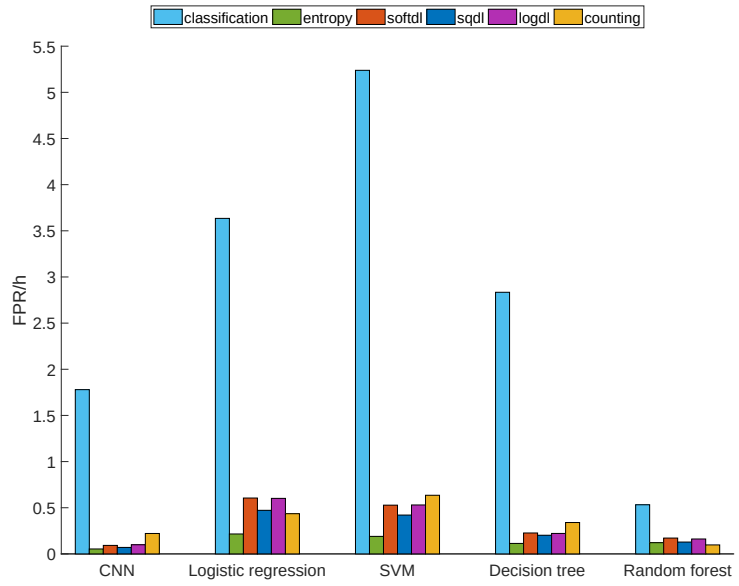
Results and discussions

In what follows, epoch-based classification results performed by CNN, logistic regression, linear SVM, decision tree, and random forest reported in this experiment are labeled by `classification` in Figures 6.4 to 6.6. After ScoreNet was applied with different loss functions *i.e.*, entropy, soft-dice loss, squared-dice loss, and log-dice loss, the performances are compared with those from the prior classification step, and are tagged by `entropy`, `softdl`, `sqdl`, `logdl`. When the counting-based method was applied, we refer to this result as `counting`. In the following figures showing the performances, color bars indicate the average values, the circle markers present the median, and the vertical bars show the interquartiles.

Figures 6.4 to 6.6 show comparisons of the event-based metrics and epoch-based metrics in all seizure detection cases, including cases of applying the counting-based method and ScoreNet to improve seizure detection results from using only the classifiers. Note that no displays of the interquartile ranges appear in Figures 6.4 and 6.5 because they are too small to be clearly visualized; the median, first and third quartiles of GDR, specificity, and accuracy were all 100%, and those of FPR/h were 0. In overall, the epoch-based seizure classifiers in the prior step detected seizure events with GDR of more than 80%, obtained F_1 and sensitivity of less than 40% and 55%, and accuracy and specificity of more than 99%, but FPR varied drastically from 0.53 per hour to 5.24 per hour on average upon the classifiers. In particular, the logistic regression, linear SVM, and decision tree achieved superior GDRs, but unfortunately yielded inferior F_1 and high FPRs per hour, whereas the random forest obtained the best F_1 and FPR/h but the lowest GDR. Moreover, high GDR and F_1 of 93.74% and 28.61%, and an intermediate FPR/h of 1.78 were obtained by the proposed CNN model. As a result, it is possible that the logistic regression, linear SVM, and decision tree randomly detected many isolated EEG epochs, resulting in high FPR/h, and low F_1 and specificity. On the other hand, the random forest possibly indicates some seizure occurrences with small seizure probabilities so that those epochs are finally decided to be normal. In addition, even though the proposed CNN is likely to produce a moderate number of false positives, the model still correctly detects both seizure events and seizure epochs. As a result, the proposed CNN without any hand-crafted features is comparable to the random forest and superior to the other classifiers with the selected significant features if we consider F_1, GDR and FPR/h as performance metrics.



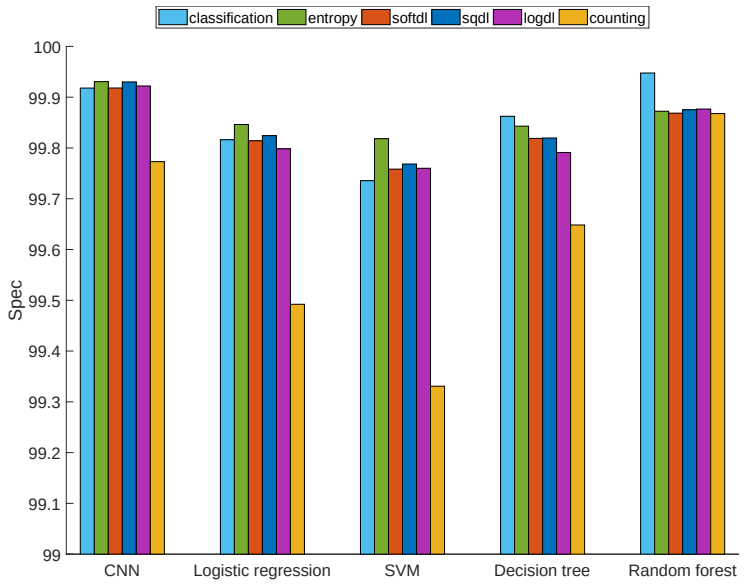
(a) GDR.



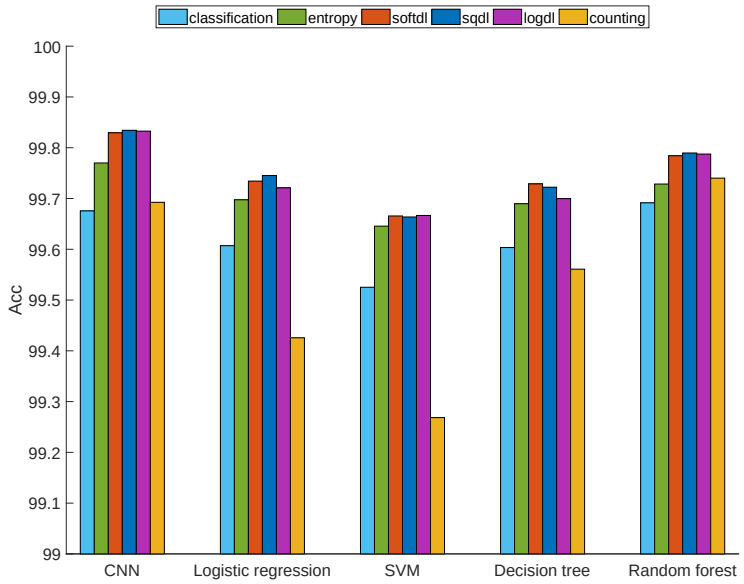
(b) FPR/h.

Figure 6.4: Comparisons of averaged GDR and FPR/h obtained from test cases using different epoch-based seizure detection methods.

On average, after applying the counting-based method and ScoreNet, regardless of the classification methods, F_1 scores increased at least 18%, FPRs were significantly reduced to at most 0.64 times per hour, but GDRs were slightly dropped. Specifically, using ScoreNet with any dice loss functions generally provided better F_1 scores and GDRs than the counting-based method, and FPRs decreased the most when the binary cross-entropy function was used. The combination of the proposed CNN and ScoreNet and the soft-dice loss accomplished the highest F_1 of 70.15%. On the other hand, the counting-based method typically yielded a higher sensitivity but lower accuracy and specificity than ScoreNet. We found



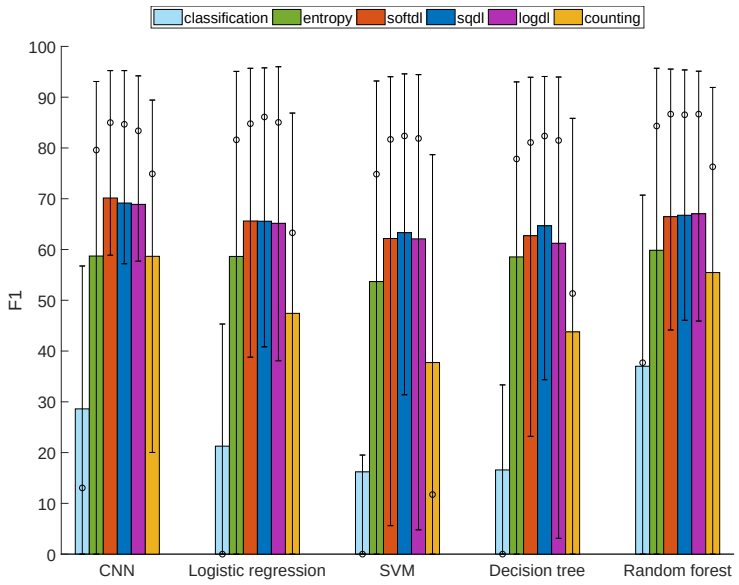
(a) Specificity.



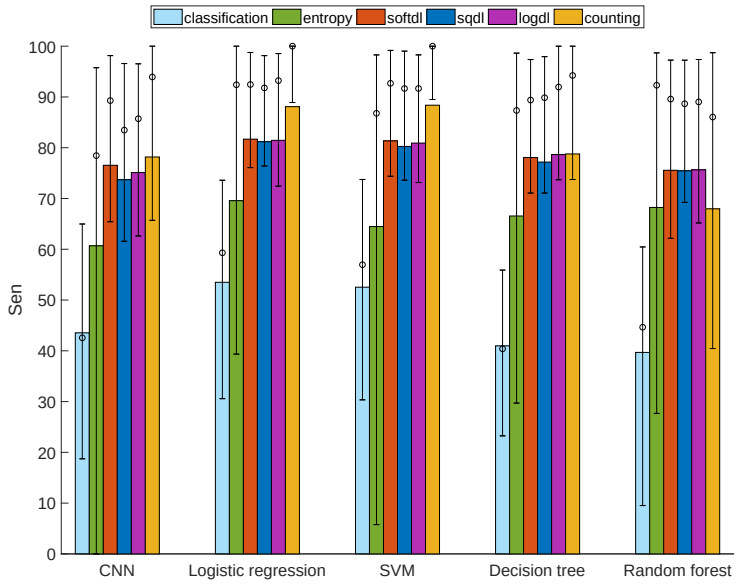
(b) Accuracy.

Figure 6.5: Comparisons of averaged specificity and accuracy obtained from test cases using different epoch-based seizure detection methods.

that the criterion (4.7) that allows seizure candidates to be seizure-detected epochs can be easily satisfied with the chosen parameters $w = 6$ and $p = 2$, so many groups of seizure candidates are eventually regarded as seizure activities. Nevertheless, many onsets and offsets of these candidate groups are detected earlier and later than the actual onsets and offsets, respectively, resulting in lower specificity and accuracy compared to that of ScoreNet. According to a large number of normal epochs, these reductions in the specificity and accuracy implies moderate numbers of the exceeding false positives. Thus, outcomes of ScoreNet with any dice loss function are more promising than those of the counting-based method.



(a) F_1.



(b) Sensitivity.

Figure 6.6: Comparisons of averaged F_1 and sensitivity obtained from test cases using different epoch-based seizure detection methods. Color bars indicate the average values, the circle markers present the median, and the vertical bars show the interquartiles.

When applying ScoreNet with the dice loss functions, we found that GDR was slightly dropped only 1 - 5%, and GDR significantly decreased about 15 - 20% when the cross-entropy function was used. On the other hand, GDR obtained from the case of the random forest increased up to 7% by combinations of ScoreNet and the dice loss functions. As shown in Figure 6.7, we found that the random forest can produce a sequence of small seizure probabilities distinguishable from the background. In this case, ScoreNet could potentially identify the seizure from the seizure probabilities. This means

that using ScoreNet generally improves epoch-based classification performance metrics. Specifically, a combination of ScoreNet and any dice loss function can also indicate some seizures from events that are not originally detected in the classification process. Thus, the dice-loss functions, including the log-dice loss, are appropriate for handling an imbalanced-class data problem.

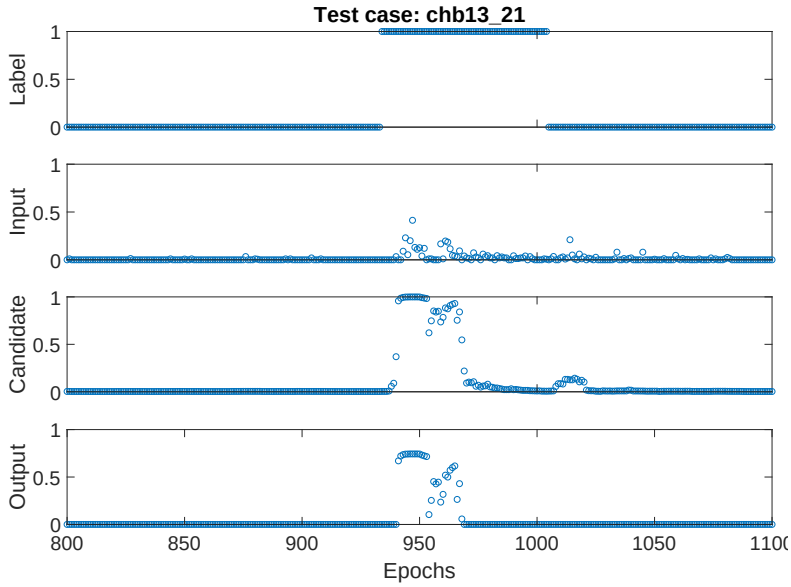


Figure 6.7: Results of ScoreNet when the log-dice loss and the random forest are used, and the test case is `chb13_21`.

From the reports of F_1, GDR, and FPR/h, we can generally interpret that some isolated predictions, *i.e.*, true positives, false positives, and false negatives, are eliminated by using the proposed seizure onset and offset determination methods, and the result of eliminating these predictions depends on the employed cost function. According to a large reduction of FPR/h, this indicates that using the cross-entropy with ScoreNet potentially reduces several isolated false positives. However, since using the cross-entropy has a bias towards the normal class, and only a few epochs are correctly classified as seizures by epoch-based classifiers, the number of predicted positives is not sufficiently high and, hence, those predicted positives are suppressed in the onset and offset detection process, resulting in a decrease in the true positives. In the case of the dice loss functions, ScoreNet generally yields similar results of seizure detection performance across any seizure detection methods. Specifically, a combination of ScoreNet and the soft-dice loss generally provides the best results of detecting seizure activities, and ScoreNet with the squared-dice loss can better reduce FPR/h. Since only a few segments during some seizure episodes were detected, the model recognized them as artifacts. As a result, GDR also was relatively lower compared to that of the other loss functions. In addition, using the log-dice loss can better improve the classification performance when classification errors are large. For instance, the results of using random forest and log-dice loss are illustrated in Figure 6.7. Seizure probabilities from the random forest (input) during the actual seizure are not sufficiently high to be detected, but they are visually distinguishable from the background. We can see that these predictions are then jointly boosted by ScoreNet so that they are potential enough to represent the seizure episode. GDR and F_1 are, thus, improved the most when using the log-dice loss in the case of random forest.

Finally, Table 6.5 summarizes a comparison of seizure detection performance noted with an amount of data and a validation scheme used in each study; the method achieving the best F_1 is chosen to compare the performance. It was found that several studies specifically selected records from the

database, and a few applied LOOCV as a validation scheme. Work in [105] used only records containing seizure in the experiment, and neither data specification nor validation scheme was reported in [68]. Moreover, some studies performed data sampling approaches to create balanced training data [45],but in practice some EEG characteristics – such as rarity and types – cannot be accurately selected; it is more clinically challenging to use all data and apply LOOCV to verify the detection performance. With these considerations in mind, our proposed method yielded competitive performances against previous results.

Table 6.5: Comparison of seizure detection methods using CHB-MIT database.

Study	Detection type	Data specification	Validation	Method	Acc	Sen	Spec	F_1	GDR	FPR/h
[56]	Event	Long seizures	No CV	aEEG + adaptive threshold	NR	NR	NR	NR	88.50	0.18
[68]	Event	NR	NR	Spectrogram + mSSDA	93.82	NR	NR	96.05	NR	NR
[57]	Event	Specific records (22 cases)	LOOCV	Energy and fractal dimension + RVM	NR	NR	NR	NR	97.00	0.10
[81]	Event	166 mins (11 cases) 70% training data	No CV	PSD + CNN-based ensemble	92.60	92.30	97.00	NR	NR	NR
[77]	Event	All Balanced training data	No CV	Statistical features + Bi-LSTM	92.66	93.61	91.85	NR	NR	NR
[86]*	Onset	NR	LOOCV	Energy + RBF SVM	NR	NR	NR	NR	96.00	0.08
[105]	Onset	Seizure in record	LOOCV	Unified multi-level spectral-temporal feature + RBF SVM	97.80	NR	NR	78.00	97.20	0.64
[85]	Onset/offset	Bipolar montage (18 cases) Balanced training data	10-fold	Relative band energy + LDA	NR	NR	99.99	NR	92.60	0.30
[89]	Onset/offset	397 hrs (18 cases) 60% training data	5-fold	Statistical features + LDA	98.00	NR	98.05	NR	100.00	4.02
[17]	Onset/offset	All	LOOCV	CNN + counting-based method	99.72	72.78	99.82	64.40	83.41	0.12
Proposed method	Onset/offset	All	LOOCV	CNN+ ScoreNet (soft-dice loss)	99.83	76.54	99.92	70.15	91.96	0.09

NR = no report, All = use full data set, * Use median instead of mean in the report

6.3 Seizure onset and offset detection

This experiment is conducted to ensure that the counting-based method and ScoreNet have abilities to indicate the seizure onset and offset of predicted seizure episodes declared by the prior classification. As shown in Section 6.2, predictions from any epoch-based seizure detector are not sufficient to be directly exploited to infer the onset and offset of the seizure. Therefore, in this experiment, results from the epoch-based seizure detector cannot be used for comparisons, and only the outcomes of the proposed seizure onset and offset detection models are compared. In addition, an *EL-index* discussed in Section 4.4 is mainly used to as a time-based metric, and a mean latency is exploited for only describing late or early prediction.

Experimental setup

This experiment was set the same as in Section 6.2. To assess the performance of seizure onset and offset detection, we exploited the usage of the EL-index explained in Section 4.4 and compared it with other latency indices. Here, the seizure onset and offset were identified by the first and last epochs of a predicted seizure episode. If more than one predicted seizure event appears during one actual event, as shown in Figure 6.8, the onset of the *first* predicted event and the offset of the *last* positive activity were used to calculate the time-based metrics. Since a time gap between any two actual seizures is practically substantial, it is unlikely that our predicted seizure could overlap with two actual events, so we omit the calculation of this case. Finally, the mean latency, mean absolute latency, and EL-index calculated from the latencies for each test case and each seizure onset and offset detector were reported and compared.

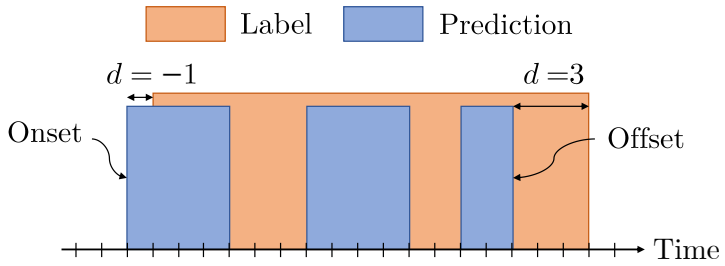


Figure 6.8: Example of determining onset and offset of seizure. Negative latency $(d < 0)$ and positive latency $(d > 0)$ mean early and late detection, respectively.

Results and discussion

As in Section 6.2, epoch-based classification results performed by CNN, logistic regression, linear SVM, decision tree, and random forest reported in this experiment are labeled by `classification` in Figures 6.10 and 6.11. After ScoreNet was applied with different loss functions *i.e.*, entropy, soft-dice loss, squared-dice loss, and log-dice loss, the performances are compared with those from the prior classification step, and are tagged by `entropy`, `softdl`, `sqdl`, `logdl`. When the counting-based method was applied, we refer to this result as `counting`. Color bars indicate the average values, the circle markers present the median, and the vertical bars show the interquartiles.

GDR, $|d|$, and the corresponding EL-indices collected from the test results are displayed in Figure 6.9, for each value of GDR shown in different colors. When GDR $= 0$, we set d to zero for the purpose of visualization in the plot. The test cases of 0% GDR shown in yellow markers imply that there is a portion of undetected events. Hence, if we use only the mean absolute latency index, these detection failures are ignored whereas EL-index can capture this since zero GDR is mapped to zero score of EL-index. As shown in Figure 6.9, at low GDR (about 40 - 50%), it is possible that seizure events were randomly detected but low mean absolute latency could be obtained whereas the EL-index is dropped (indicating worse performance). Therefore, the proposed EL-index is more suitable for being a time-based index than the mean absolute latency.

For the cases of non-zero GDR, the EL-index is dominantly high when the detection delay is insignificant and many focused events are detected. From Figure 6.9, the relationship of EL-index and mean absolute latency mostly satisfies the exponential bound, GDR $\cdot\, r^{|d|}$ represented in dashed lines. As analyzed in Section 4.4, this means that latencies from detecting seizure onsets/offsets mostly have low variation. In addition, consider markers above the dashed line for a specific GDR. The mean absolute latency definitely cannot differentiate cases of similar and different onset/offset latencies, whereas the EL-index does. Therefore, the EL-index provides not only a meaning of accurate seizure onset/offset detection but also the accuracy of seizure event detection and the interpretation of latency distributions when being considered jointly with GDR.

Figure 6.10 illustrates mean onset and offset latencies from each detection method to show trends of the detection errors. We discovered that using ScoreNet yielded positive medians and averages of mean onset latencies, and those of mean offset latencies were negative. On the other hand, the counting-based method gave negative medians of the mean onset latencies and positive medians of the mean offset latencies, but the averages of those latencies varied depending on the epoch-based classifier. In the case of ScoreNet, the magnitudes of the positive mean onset latencies seem to be higher than those of the negative ones, whereas the magnitudes of the negative mean offset latencies are larger than those of the positive ones. According to the characteristics of seizures, patterns of pre-ictal and post-ictal states, state before and after the actual seizure, dominantly appear less than ictal patterns. In practice, it is highly possible that neurologists annotate the starting and ending points of the seizure at the pre-ictal and post-ictal states to ensure that the annotation fully covers the event. Hence, ScoreNet generally indicates a seizure episode smaller than the actual event. For the counting-based method, the outcomes

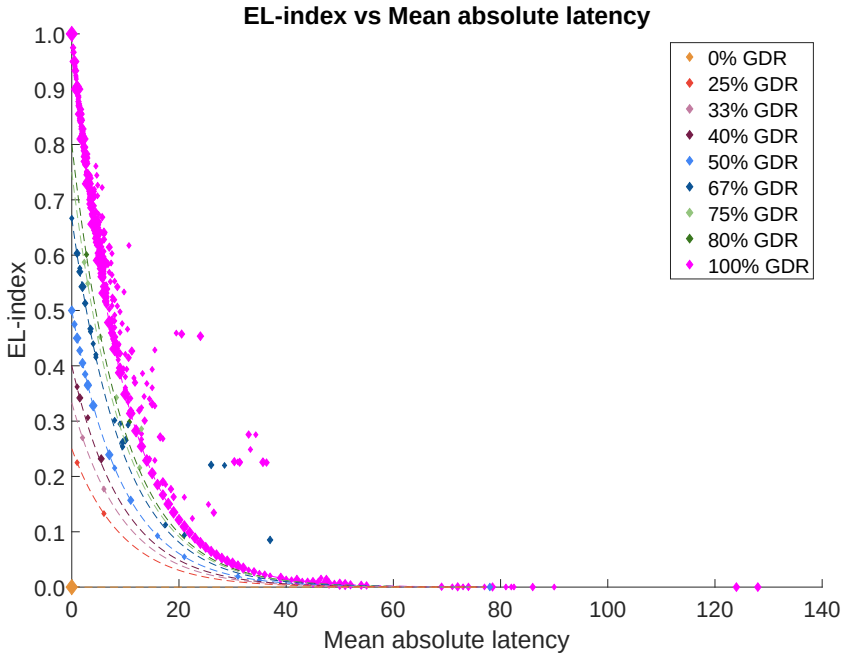
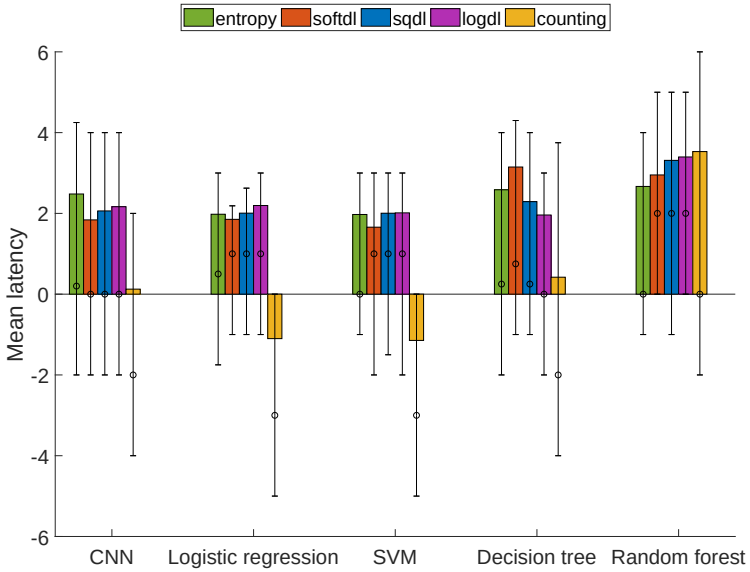


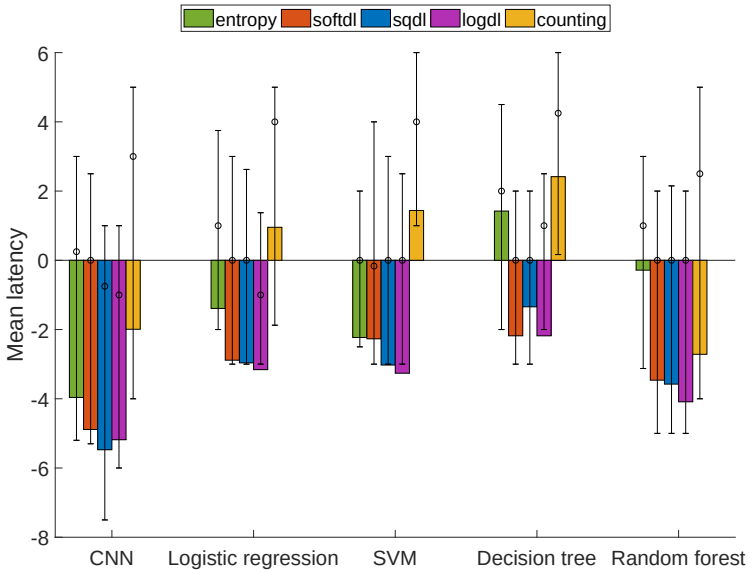
Figure 6.9: Relation of EL-index and average of absolute latency from test data given $r = 0.9$. The marker size is proportional to the number of samples in a log scale, and a dashed line illustrates GDR $\cdot\, r^{|d|}$. We set $d = 0$ when GDR $= 0$ for the purpose of visualization.

tend to be contrary to those of ScoreNet. According to the medians, the mean onset latencies tend to be negative, *i.e.*, early detection, and the mean offset latencies are likely to be positive, *i.e.*, late detection. As discussed in Section 6.2, there are predicted seizures longer than actual events. In this case, the false positives occurring before the annotated seizure onset and after the labeled seizure offset resulted in early and late predictions of the onset and the offset, respectively. As a result, the counting-based method seems to indicate a seizure event that excessively covers an actual seizure.

Figure 6.11 shows comparisons of EL-indices of detecting seizure onsets and offsets using various methods when $r = 0.9$, and Table 6.6 concludes EL-indices obtained from all methods. In the case of seizure onset detection, the mean values of EL-indices ranged from 0.50 to 0.71, and the medians ranged from 0.59 to 0.81. For detecting seizure offset, means and medians of EL-indices ranged from 0.45 to 0.67 and from 0.53 to 0.73, respectively. As shown in Figure 6.9 and from the minimum median of 0.53, this implies that seizure onsets and offsets are typically detected with mean absolute errors less than 10 seconds, which is clinically acceptable. In particular, EL-indices of indicating seizure onset and offset using ScoreNet with dice loss functions were similarly high compared to those of the other methods, and applying the log-dice loss with ScoreNet achieved slightly better EL-indices of detecting seizure offset among the other dice loss functions. It is evident that all seizure onset and offset detection methods could better indicate seizure onsets than seizure offsets. This is due to epileptic seizure characteristics that ictal patterns establish at the end of an event less dominantly than at the beginning, so predictions of seizure epochs tend to appear less at the event termination. This also means that ScoreNet with the log-dice loss can better fix incorrect classification outcomes having low seizure probabilities at the ending of the event. Thus, exploiting the log-dice loss is more useful when a large margin of detection error occurs.

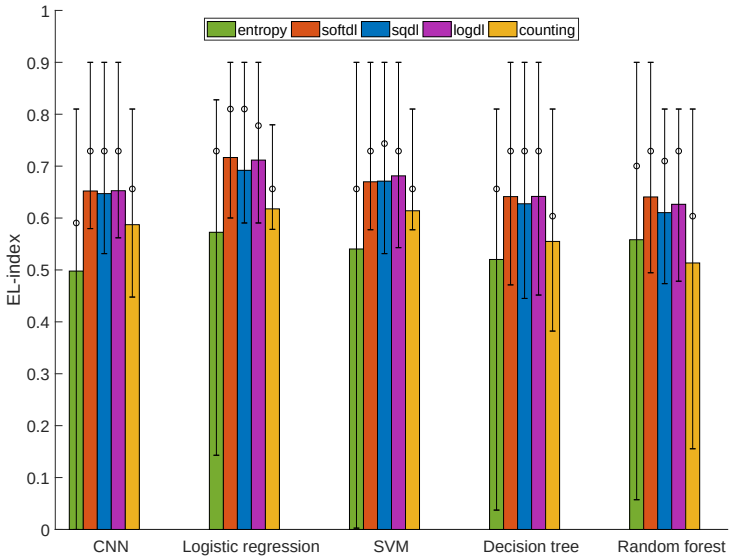


(a) Seizure onset.

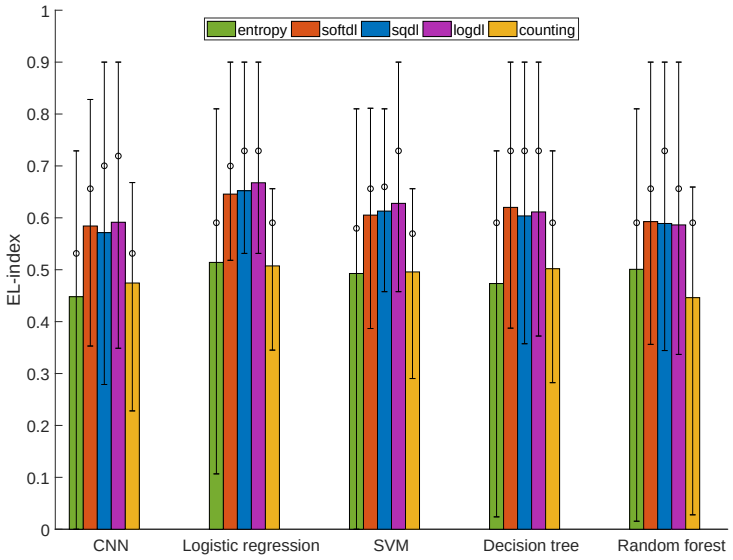


(b) Seizure offset.

Figure 6.10: Comparisons of mean latencies of detecting seizure onset and offset.



(a) Seizure onset.



(b) Seizure offset.

Figure 6.11: Comparisons of EL-indices of detecting seizure onset and offset.

Table 6.6: Summary of mean EL-index of seizure onset and offset determination. The maximum EL-index from each classifier is given in boldface.

		CNN	Logistic regression	SVM	Decision tree	Random forest
Onset	entropy	0.50	0.57	0.54	0.52	0.56
	softdl	**0.65**	**0.72**	0.67	**0.64**	**0.64**
	sqdl	**0.65**	0.69	0.67	0.63	0.61
	logdl	**0.65**	0.71	**0.68**	**0.64**	0.63
	counting	0.59	0.62	0.61	0.55	0.51
Offset	entropy	0.45	0.51	0.49	0.47	0.50
	softdl	0.58	0.65	0.61	**0.62**	**0.59**
	sqdl	0.57	0.65	0.61	0.60	**0.59**
	logdl	**0.59**	**0.67**	**0.63**	0.61	**0.59**
	counting	0.47	0.51	0.50	0.50	0.45

Chapter VII

CONCLUSION AND FUTURE WORK

This chapter concludes the dissertation with goals and contributions followed by limitations and future work.

7.1 Goals and contributions

This dissertation aimed to develop and validate a method of seizure onset and offset detection. The contributions of this dissertation are as follows.

In Section 4.1, we presented a way to evaluate an individual feature by using the Bayesian error and non-parametric probability distribution estimation. Features from time, frequency, and time-frequency domains were used to determine the significance and the redundancy. Two records containing at least one seizure from every case of patient in the CHB-MIT Scalp EEG database were randomly selected to conduct the experiment. As reported in Section 6.1, features related to changes in amplitude, frequency, and rhythmicity of EEGs were significant to distinguish ictal patterns from normal EEGs. In particular, energy, variance, nonlinear energy, and Shannon entropy extracted from raw EEGs, and energy and variance computed from DWT coefficients accomplished the highest improvement rates, so we recommend using these features in this application.

We showed how to design a CNN model of which the input was a raw EEG segment to detect an epileptic seizure episode in Section 4.2. Prior information of electrode placement was exploited to construct the model input and to design the model structure. Comparisons of the proposed model with other epoch-based classifiers were demonstrated using the same data set. As shown in Section 6.2, the CNN model without any specific features, achieving F_1 of 28.61%, GDR of 93.74%, and FPR/h of 1.78 was competitive with the random forest and outperformed the other classifiers. The performances of low FPR/h and high GDR accomplished by the proposed model are clinically acceptable since the model could detect almost all seizures event with an immediate number of false alarms per hour.

From Section 4.3, a counting-based method and ScoreNet were proposed to determine the seizure onset and offset from detected seizure episodes. A condition of no abrupt change in EEG was used to design the counting-based method, and ScoreNet was established by adding flexibility to the counting-based method. We also introduced a log-dice loss based on a dice similarity coefficient to handle an imbalanced data problem. Additionally, an EL-index was presented to measure the precision of seizure onset and offset detection. As shown in Section 6.3, these methods could typically determine the seizure onset and offset with an error of less than 10 seconds. Moreover, results in Section 6.2 showed that these methods could generally reduce many false alarms and significantly improve F_1 from the prior results compromising with a slightly lower GDR. The best F_1 of 70.15% was obtained by the combination of the CNN model and ScoreNet with a soft-dice loss. According to nonlinear penalty, the log-dice loss could better increase the detection performances than other losses when predictions from the prior classification stage were largely incorrect. In addition, we discovered that the EL-index is suitable for measuring a detection delay and also distinguishes different empirical distributions of latencies when used with GDR and a mean absolute latency. Finally, ScoreNet has an ability to boost up performances from a potential classifier where the improvement level is varied upon the classifier, and it can be easily extended to multi-class classification problem by changing the dimensions of ScoreNet parameters.

7.2 Limitations and future work

This study can be extended in several directions, and we include some interesting topics below.

Universal detector. As this dissertation considers a patient-specific scheme, data used to train and validate a method must be collected from the same patient. Since the developed method needs to be early trained before tested, it means that we need to record EEGs of the patient first. This becomes a limitation of the method in some practical conditions that patients cannot be monitored beforehand. Future work can focus on a patient non-specific epoch-based seizure detector so that the data for training can be universally obtained. However, the performance of a universal detector can be lower on average since some characteristics of seizures partially depend on profiles of each patient. To address this issue, more highly discriminative features representing general seizure characteristics should be explored. For instance, estimated seizure areas and locations are informative to the detection as these provide spatial characteristics of seizures. This spatial information implicitly obtained by using independent component analysis (ICA) [106] can be exploited as prior information for detecting the seizures.

Real-time detection. In Section 4.3, the counting-based method and ScoreNet require all epoch-based detection results from long EEG signals to be collected offline. The main issue for this implementation occurs at a process of grouping seizure candidates since data must be collected to form candidate groups. This means that these methods cannot be readily implemented for real-time detection since clinical EEG data are continuously streamed. Therefore, future research will focus on modifying their algorithms and applying these models to the streaming data. The implementation should be modified in a way that output gates and outputs can be iteratively calculated once a new candidate group is formed.

Multi-modality. As explained in Section 1.2, we limited our study to only one biosignal, EEG, to automatically detect epileptic seizures. However, in clinic, affected patients are usually observed by experts using a combination of multiple modalities such as continuous video-EEG monitoring. Moreover, it was found that using ECG as another biological source reveals more information when EEG alone is unreliable [54]. Therefore, in the future, advantages of jointly applying multiple tools such as ECG, EMG, and video of the monitoring can be examined. Data fusion approaches, *e.g.*, independent vector analysis (IVA) [107], can be used to determine connections such as synchronous activities from those modalities. Moreover, as generalization of ICA, IVA can also reveal common information of seizures from different patient demographics.

Analysis of data usage. In this dissertation, we mainly aimed to develop an automatic detection model of epileptic seizure onsets and offsets, and partially discussed that an imbalance ratio is a key factor to improved performances reported in this dissertation. Experimental results in Appendix C.5 confirmed our hypothesis that there were strong correlations between the detection accuracy and the imbalance ratio in most test cases of several classification methods. However, given an amount of training data and a ratio of abnormal to normal classes, we could not analytically specify a lower bound of detection performance, and we are not certain if such analysis can be feasibly performed. This is due to the nature of neural networks or other machine learning approaches that generally involve solving optimization problems where the estimated model parameters are highly nonlinear and implicit functions of EEG record length, the number of epochs, and the imbalance ratio, let alone the resulting detection performances. From a practical point of view, we agree that it is beneficial to ensure the detection performances in some degree for a given set of training data because this is useful for neurologists to only focus on labeling the particular set of data; however, to the best of our knowledge, we believe that this can be done empirically at this stage. A future study can systematically explore requirements of the training set such as types of seizures, an appropriate imbalance ratio, and the minimum amount of the data. As an example, to determine the minimum data size and imbalance ratio, we can increase or duplicate seizure-class

samples, or decrease normal class samples. In addition, since the annotation of the CHB-MIT Scalp EEG database does not contain types of seizures, collaboration with neurologists can be established to acquire labels of the seizure types.

REFERENCES

[1] R. Fisher, C. Acevedo, A. Arzimanoglou, A. Bogacz, J. Cross, C. Elger, J. J. Engel, L. Forsgren, J. French, M. Glynn et al., "ILAE official report: a practical clinical definition of epilepsy," Epilepsia, vol. 55, no. 4, pp. 475–482, 2014.

[2] D. Thurman, E. Beghi, C. Begley, A. Berg, J. Buchhalter, D. Ding, D. Hesdorffer, W. Hauser, L. Kazis, R. Kobau et al., "Standards for epidemiologic studies and surveillance of epilepsy," Epilepsia, vol. 52, no. 7, pp. 2–26, 2011.

[3] A. Blum and S. Rutkove, The Clinical Neurophysiology Primer. Springer, 2007, vol. 388.

[4] A. Rowan and E. Tolunsky, A primer of EEG: with a mini-atlas. Butterworth-Heinemann Medical, 2003.

[5] S. Schachter and D. Schomer, The Comprehensive Evaluation and Treatment of Epilepsy: a practical guide. Elsevier, 1997.

[6] M. Mastrangelo, I. Fiocchi, P. Fontana, G. Gorgone, G. Lista, and V. Belcastro, "Acute neonatal encephalopathy and seizures recurrence: a combined aEEG/EEG study," Seizure, vol. 22, no. 9, pp. 703–707, 2013.

[7] V. Sridevi, M. Reddy, K. Srinivasan, K. Radhakrishnan, C. Rathore, and D. Nayak, "Improved patient-independent system for detection of electrical onset of seizures," Journal of Clinical Neurophysiology, vol. 36, no. 1, pp. 14–24, 2019.

[8] R. Fisher, J. Cross, C. D'souza, J. French, S. Haut, N. Higurashi, E. Hirsch, F. Jansen, L. Lagae, S. Moshé et al., "Instruction manual for the ILAE 2017 operational classification of seizure types," Epilepsia, vol. 58, no. 4, pp. 531–542, 2017.

[9] R. Nardou, D. Ferrari, and Y. Ben-Ari, "Mechanisms and effects of seizures in the immature brain," in Seminars in Fetal and Neonatal Medicine, vol. 18, no. 4. Elsevier, 2013, pp. 175–184.

[10] B. B. Ma and V. R. Rao, "Responsive neurostimulation: candidates and considerations," Epilepsy & Behavior, 2018.

[11] E. B. Geller, "Responsive neurostimulation: Review of clinical trials and insights into focal epilepsy," Epilepsy & Behavior, 2018.

[12] K. Vonck, R. Raedt, and P. Boon, "Vagus nerve stimulation and the postictal state," Epilepsy & Behavior, vol. 19, no. 2, pp. 182–185, 2010.

[13] M. Goldenberg, "Overview of drugs used for epilepsy and seizures: etiology, diagnosis, and treatment," Pharmacy and Therapeutics, vol. 35, no. 7, pp. 392–415, 2010.

[14] N. So and W. Blume, "The postictal EEG," Epilepsy & Behavior, vol. 19, no. 2, pp. 121–126, 2010.

[15] P. Boonyakitanont, A. Lek-uthai, K. Chomtho, and J. Songsiri, "A comparison of deep neural networks for seizure detection in EEG signals," bioRxiv preprint, p. 702654, 2019, https://www.biorxiv.org/content/10.1101/702654v1.

[16] ——, "A review of feature extraction and performance evaluation in epileptic seizure detection using EEG," Biomedical Signal Processing and Control, vol. 57, p. 101702, 2020.

[17] P. Boonyakitanont, A. Lek-uthai, and J. Songsiri, "Automatic epileptic seizure onset-offset detection based on CNN in scalp EEG," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 1225–1229.

[18] P. Boonyakitanont, A. Lek-uthai, and J. Songsiri, "ScoreNet: A neural network-based post-processing model for identifying epileptic seizure onset and offset in EEGs," bioRxiv preprint, 2020, https://www.biorxiv.org/content/10.1101/2020.12.21.423728v2.

[19] W. Blume, G. Young, and J. Lemieux, "EEG morphology of partial epileptic seizures," Electroencephalography and Clinical Neurophysiology, vol. 57, no. 4, pp. 295–302, 1984.

[20] F. Pauri, F. Pierelli, G. Chatrian, and W. Erdly, "Long-term EEG-video-audio monitoring: computer detection of focal EEG seizure patterns," Electroencephalography and Clinical Neurophysiology, vol. 82, no. 1, pp. 1–9, 1992.

[21] V. Vapnik, "Principles of risk minimization for learning theory," in Advances in Neural Information Processing Systems, 1992, pp. 831–838.

[22] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media, 2009.

[23] C. M. Bishop, Pattern recognition and machine learning. Springer, 2006.

[24] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.

[27] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, M. Hasan, B. C. Van Essen, A. A. Awwal, and V. K. Asari, "A state-of-the-art survey on deep learning theory and architectures," Electronics, vol. 8, no. 3, pp. 292–358, 2019.

[28] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 448–456.

[30] B. Xu, R. Huang, and M. Li, "Revise saturated activation functions," arXiv preprint arXiv:1602.05980, 2016.

[31] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Proceedings of the 27th international conference on machine learning, 2010, pp. 807–814.

[32] R. Lguensat, M. Sun, R. Fablet, P. Tandeo, E. Mason, and G. Chen, "EddyNet: A deep neural network for pixel-wise classification of oceanic eddies," in Proceedings of the 2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2018, pp. 1764–1767.

[33] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in Proceedings of the 2016 International Conference on 3D Vision (3DV). IEEE, 2016, pp. 565–571.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[35] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, and H. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet," Circulation, vol. 101, no. 23, pp. e215–e220, 2000.

[36] U. Acharya, S. Sree, G. Swapna, R. Martis, and J. Suri, "Automated EEG analysis of epilepsy: A review," Knowledge-Based Systems, vol. 45, pp. 147–165, 2013.

[37] T. Alotaiby, F. El-Samie, S. Alshebeili, and I. Ahmad, "A review of channel selection algorithms for EEG signal processing," EURASIP Journal on Advances in Signal Processing, vol. 2015, no. 1, pp. 66–86, 2015.

[38] E. Alickovic, J. Kevric, and A. 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, vol. 39, pp. 94–102, 2018.

[39] A. Aarabi, F. Wallois, and R. Grebe, "Automated neonatal seizure detection: a multistage classification system through feature selection based on relevance and redundancy analysis," Clinical Neurophysiology, vol. 117, no. 2, pp. 328–340, 2006.

[40] M. Ahmad, M. Saeed, S. Saleem, and A. Kamboh, "Seizure detection using EEG: A survey of different techniques," in Proceedings of the 2016 International Conference on Emerging Technologies. IEEE, 2016, pp. 1–6.

[41] J. Gotman, "Automatic recognition of epileptic seizures in the EEG," Electroencephalography and Clinical Neurophysiology, vol. 54, no. 5, pp. 530–540, 1982.

[42] L. Guo, D. Rivero, J. Dorado, J. Rabunal, and A. Pazos, "Automatic epileptic seizure detection in EEGs based on line length feature and artificial neural networks," Journal of Neuroscience Methods, vol. 191, no. 1, pp. 101–109, 2010.

[43] L. Orosco, E. Laciar, A. Correa, A. Torres, and J. Graffigna, "An epileptic seizures detection algorithm based on the empirical mode decomposition of EEG," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2009, pp. 2651–2654.

[44] V. Gupta and R. B. Pachori, "Epileptic seizure identification using entropy of FBSE based EEG rhythms," Biomedical Signal Processing and Control, vol. 53, p. 101569, 2019.

[45] A. R. Hassan, A. Subasi, and Y. Zhang, "Epilepsy seizure detection using complete ensemble empirical mode decomposition with adaptive noise," Knowledge-Based Systems, vol. 191, p. 105333, 2020.

[46] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "EEG-based neonatal seizure detection with support vector machines," Clinical Neurophysiology, vol. 122, no. 3, pp. 464–473, 2011.

[47] T. Kumar, V. Kanhangad, and R. Pachori, "Classification of seizure and seizure-free EEG signals using local binary patterns," Biomedical Signal Processing and Control, vol. 15, pp. 33–40, 2015.

[48] A. Jaiswal and H. Banka, "Local pattern transformation based feature extraction techniques for classification of epileptic EEG signals," Biomedical Signal Processing and Control, vol. 34, pp. 81–92, 2017.

[49] U. Acharya, F. Molinari, S. Sree, S. Chattopadhyay, K. Ng, and J. Suri, "Automated diagnosis of epileptic EEG using entropies," Biomedical Signal Processing and Control, vol. 7, no. 4, pp. 401–408, 2012.

[50] U. Acharya, H. Fujita, V. Sudarshan, S. Bhat, and J. Koh, "Application of entropies for automated diagnosis of epilepsy using EEG signals: A review," Knowledge-Based Systems, vol. 88, pp. 85–96, 2015.

[51] J. Li, J. Yan, X. Liu, and G. Ouyang, "Using permutation entropy to measure the changes in EEG signals during absence seizures," Entropy, vol. 16, no. 6, pp. 3049–3061, 2014.

[52] N. Tawfik, S. Youssef, and M. Kholief, "A hybrid automated detection of epileptic seizures in EEG records," Computers & Electrical Engineering, vol. 53, pp. 177–190, 2016.

[53] R. Agarwal and J. Gotman, "Adaptive segmentation of electroencephalographic data using a non-linear energy operator," in Proceedings of the 1999 IEEE International Symposium on Circuits and Systems, vol. 4. IEEE, 1999, pp. 199–202.

[54] A. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[55] A. G. Correa, L. Orosco, P. Diez, and E. Laciar, "Automatic detection of epileptic seizures in long-term EEG records," Computers in Biology and Medicine, vol. 57, pp. 66–73, 2015.

[56] C. Satirasethawong, A. Lek-Uthai, and K. Chomtho, "Amplitude-integrated EEG processing and its performance for automatic seizure detection," in Proceedings of the 2015 IEEE International Conference on Signal and Image Processing Applications. IEEE, 2015, pp. 551–556.

[57] L. Vidyaratne and K. Iftekharuddin, "Real-time epileptic seizure detection using EEG," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 11, pp. 2146–2156, 2017.

[58] V. Bajaj and R. B. Pachori, "Classification of seizure and nonseizure EEG signals using empirical mode decomposition," IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 6, pp. 1135–1142, 2011.

[59] P. Li, C. Karmakar, J. Yearwood, S. Venkatesh, M. Palaniswami, and C. Liu, "Detection of epileptic seizure based on entropy analysis of short-term EEG," PLOS ONE, vol. 13, no. 3, p. e0193691, 2018.

[60] K. Polat and S. Güneş, "Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform," Applied Mathematics and Computation, vol. 187, no. 2, pp. 1017–1026, 2007.

[61] O. Faust, U. Acharya, L. Min, and B. Sputh, "Automatic identification of epileptic and background EEG signals using frequency domain parameters," International Journal of Neural Systems, vol. 20, no. 02, pp. 159–176, 2010.

[62] A. Subasi and M. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," Expert Systems with Applications, vol. 37, no. 12, pp. 8659–8666, 2010.

[63] L. Guo, D. Rivero, and A. Pazos, "Epileptic seizure detection using multiwavelet transform based approximate entropy and artificial neural networks," Journal of Neuroscience Methods, vol. 193, no. 1, pp. 156–163, 2010.

[64] K. Samiee, P. Kovacs, and M. Gabbouj, "Epileptic seizure classification of EEG time-series using rational discrete short-time Fourier transform," IEEE Transactions on Biomedical Engineering, vol. 62, no. 2, pp. 541–552, 2015.

[65] S. Anand, S. Jaiswal, and P. Ghosh, "Automatic focal epileptic seizure detection in EEG signals," in Proceedings of the 2017 IEEE International WIE Conference on Electrical and Computer Engineering. IEEE, 2017, pp. 103–107.

[66] D. Wang, D. Miao, and C. Xie, "Best basis-based wavelet packet entropy feature extraction and hierarchical EEG classification for epileptic detection," Expert Systems with Applications, vol. 38, no. 11, pp. 14 314–14 320, 2011.

[67] S. Janjarasjitt, "Epileptic seizure classifications of single-channel scalp EEG data using wavelet-based features and svm," Medical & Biological Engineering & Computing, vol. 55, no. 10, pp. 1743–1761, 2017.

[68] Y. Yuan, G. Xun, K. Jia, and A. Zhang, "A multi-view deep learning method for epileptic seizure detection using short-time fourier transform," in Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM, 2017, pp. 213–222.

[69] S. Ammar and M. Senouci, "Seizure detection with single-channel EEG using extreme learning machine," in Proceedings of the 17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering. IEEE, 2016, pp. 776–779.

[70] P. Fergus, A. Hussain, D. Hignett, D. Al-Jumeily, K. Abdel-Aziz, and H. Hamdan, "A machine learning system for automated whole-brain seizure detection," Applied Computing and Informatics, vol. 12, no. 1, pp. 70–89, 2016.

[71] R. San-Segundo, M. Gil-Martín, L. F. D'Haro-Enríquez, and J. M. Pardo, "Classification of epileptic EEG recordings using signal transforms and convolutional neural networks," Computers in Biology and Medicine, vol. 109, pp. 148–158, 2019.

[72] A. Shoeibi, N. Ghassemi, M. Khodatars, M. Jafari, S. Hussain, R. Alizadehsani, P. Moridian, A. Khosravi, H. Hosseini-Nejad, M. Rouhani et al., "Epileptic seizure detection using deep learning techniques: A review," arXiv preprint arXiv:2007.01276, 2020.

[73] H. Takahashi, A. Emami, T. Shinozaki, N. Kunii, T. Matsuo, and K. Kawai, "Convolutional neural network with autoencoder-assisted multiclass labelling for seizure detection based on scalp electroencephalography," Computers in Biology and Medicine, p. 104016, 2020.

[74] A. Shoeibi, N. Ghassemi, R. Alizadehsani, M. Rouhani, H. Hosseini-Nejad, A. Khosravi, M. Panahiazar, and S. Nahavandi, "A comprehensive comparison of handcrafted features and convolutional autoencoders for epileptic seizures detection in EEG signals," Expert Systems with Applications, vol. 163, p. 113788, 2021.

[75] U. Acharya, S. Oh, Y. Hagiwara, J. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," Computers in Biology and Medicine, vol. 100, no. 1, pp. 270–278, 2018.

[76] X. Tian, Z. Deng, W. Ying, K. Choi, D. Wu, B. Qin, J. Wang, H. Shen, and S. Wang, "Deep multi-view feature learning for EEG-based epileptic seizure detection," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 27, no. 10, pp. 1962–1972, 2019.

[77] X. Hu, S. Yuan, F. Xu, Y. Leng, K. Yuan, and Q. Yuan, "Scalp EEG classification using deep Bi-LSTM network for seizure detection," Computers in Biology and Medicine, vol. 124, p. 103919, 2020.

[78] A. O'Shea, G. Lightbody, G. Boylan, and A. Temko, "Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture," Neural Networks, vol. 123, pp. 12–25, 2020.

[79] T. Sakai, T. Shoji, N. Yoshida, K. Fukumori, Y. Tanaka, and T. Tanaka, "Scalpnet: Detection of spatiotemporal abnormal intervals in epileptic EEG using convolutional neural networks," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 1244–1248.

[80] S. Chakraborti, A. Choudhary, A. Singh, R. Kumar, and A. Swetapadma, "A machine learning based method to detect epilepsy," International Journal of Information Technology, pp. 1–7, 2018.

[81] Y. Gao, B. Gao, Q. Chen, J. Liu, and Y. Zhang, "Deep convolutional neural network-based epileptic electroencephalogram (EEG) signal classification," Frontiers in Neurology, vol. 11, 2020.

[82] A. Emami, N. Kunii, T. Matsuo, T. Shinozaki, K. Kawai, and H. Takahashi, "Seizure detection by convolutional neural network-based analysis of scalp electroencephalography plot images," NeuroImage: Clinical, vol. 22, p. 101684, 2019.

[83] I. Covert, B. Krishnan, I. Najm, J. Zhan, M. Shore, J. Hixson, and M. J. Po, "Temporal graph convolutional networks for automatic seizure detection," arXiv preprint arXiv:1905.01375, 2019.

[84] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[85] L. Orosco, A. Correa, P. Diez, and E. Laciar, "Patient non-specific algorithm for seizures detection in scalp EEG," Computers in Biology and Medicine, vol. 71, pp. 128–134, 2016.

[86] A. Shoeb and J. Guttag, "Application of machine learning to epileptic seizure detection," in Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 975–982.

[87] S. Janjarasjitt, "Performance of epileptic single-channel scalp EEG classifications using single wavelet-based features," Australasian Physical & Engineering Sciences in Medicine, vol. 40, no. 1, pp. 57–67, 2017.

[88] A. Shoeb, A. Kharbouch, J. Soegaard, S. Schachter, and J. Guttag, "An algorithm for detecting seizure termination in scalp EEG," in Proceedings of 2011 the Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2011, pp. 1443–1446.

[89] G. Chandel, P. Upadhyaya, O. Farooq, and Y. Khan, "Detection of seizure event and its onset/offset using orthonormal triadic wavelet based features," IRBM, vol. 40, no. 2, pp. 103–112, 2019.

[90] H. Ocak, "Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy," Expert Systems with Applications, vol. 36, no. 2, pp. 2027–2036, 2009.

[91] A. Tzallas, M. Tsipouras, and D. Fotiadis, "Automatic seizure detection based on time-frequency analysis and artificial neural networks," Computational Intelligence and Neuroscience, vol. 2007, 2007.

[92] M. Bandarabadi, C. Teixeira, J. Rasekhi, and A. Dourado, "Epileptic seizure prediction using relative spectral power features," Clinical Neurophysiology, vol. 126, no. 2, pp. 237–248, 2015.

[93] L. Devroye, L. Györfi, and G. Lugosi, A Probabilistic Theory of Pattern Recognition. Springer Science & Business Media, 2013, vol. 31.

[94] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed. Academic Press, Inc., 1990.

[95] E. Parzen, "On estimation of a probability density function and mode," The Annals of Mathematical Statistics, vol. 33, no. 3, pp. 1065–1076, 1962.

[96] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[97] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," arXiv preprint arXiv:1212.5701, 2012.

[98] Q. V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng, "On optimization methods for deep learning," in Proceedings of the 28th International Conference on International Conference on Machine Learning, 2011, pp. 265–272.

[99] R. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. 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, vol. 64, no. 6, p. 061907, 2001.

[100] V. Shah, E. von Weltin, S. Lopez, J. R. McHugh, L. Veloso, M. Golmohammadi, I. Obeid, and J. Picone, "The temple university hospital seizure detection corpus," Frontiers in Neuroinformatics, vol. 12, p. 83, 2018.

[101] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "Performance assessment for EEG-based neonatal seizure detectors," Clinical Neurophysiology, vol. 122, no. 3, pp. 474–482, 2011.

[102] D. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Techonologies, vol. 2, pp. 37–63, 2011.

[103] M. Hall and L. Smith, "Feature subset selection: a correlation based filter approach," in Proceedings of the 4th International Conference on Neural Information Processing and Intelligent Information Systems. Springer, 1997, pp. 855–858.

[104] "Seizure Detection," https://www.persyst.com/technology/seizure-detection/, accessed: 2019-4-25.

[105] F. Tang, Y. Liu, Y. Li, and Z. Peng, "A unified multi-level spectral–temporal feature learning framework for patient-specific seizure onset detection in EEG signals," Knowledge-Based Systems, vol. 205, p. 106152, 2020.

[106] V. K. Harpale and V. K. Bairagi, "Significance of independent component analysis (ICA) for epileptic seizure detection using eeg signals," in Proceedings of the International Conference on Data Engineering and Communication Technology. Springer, 2017, pp. 829–838.

[107] T. Adali, M. Anderson, and G. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," IEEE Signal Processing Magazine, vol. 31, no. 3, pp. 18–33, 2014.

[108] N. Huang, Hilbert-Huang transform and its applications. World Scientific, 2014, vol. 16.

[109] R. Esteller, J. Echauz, T. Tcheng, B. Litt, and B. Pless, "Line length: an efficient feature for seizure onset detection," in Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 2. IEEE, 2001, pp. 1707–1710.

[110] J. Kaiser, "On a simple algorithm to calculate the energy of a signal," in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1990, pp. 381–384.

[111] C. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 1948.

[112] S. Pincus, "Approximate entropy as a measure of system complexity," Proceedings of the National Academy of Sciences, vol. 88, no. 6, pp. 2297–2301, 1991.

[113] J. Richman and J. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," American Journal of Physiology-Heart and Circulatory Physiology, vol. 278, no. 6, pp. H2039–H2049, 2000.

[114] B. Hjorth, "EEG analysis based on time domain properties," Electroencephalography and Clinical Neurophysiology, vol. 29, no. 3, pp. 306–310, 1970.

[115] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research," ASU Feature Selection Repository, pp. 1–28, 2010.

Appendix A

FEATURE EXTRACTION AND CFS

This chapter describes the details of features used as well as a correlation-based feature selection (CFS) in this dissertation.

A.1 Feature extraction

In this section, we categorize features by feature domains: time domain and frequency domain. Time-domain features are those calculated on raw EEG signals or on pre-processed signals done in the time domain, such as empirical mode decomposition [108]. On the other hand, frequency-domain features are computed on discrete-Fourier transform of raw EEG signals. In addition, time-frequency-domain features also use these descriptions and are computed from coefficients of a decomposition technique such as discrete wavelet transform and wavelet packet decomposition. To what follows, we denote $X = [x_1, x_2, \ldots, x_N]$ a sequence of length N used for extracting a feature. For instance, X can be an epoch of raw EEG segment, absolute values of a raw EEG segment, power spectral density (PSD), approximation or detail coefficients from any wavelet transform, or intrinsic mode functions from empirical mode decomposition. In this section, well-known features such as statistical parameters and energy are briefly described, and those involving an uncertainty concept such as entropy are explained with mathematical expressions.

Time-domain features

1. Groups of statistical parameters have been frequently used to discriminate between ictal and normal patterns because it is assumed that EEG statistical distributions during seizure and normal periods are different. These parameters are *mean, variance, mode, median*, *skewness* (third moment describing data asymmetry), and *kurtosis* (fourth moment determining tailedness of the distribution). The *minimum* and *maximum* values are also used to quantify the range of data or the magnitude of signal baseline. *Coefficient of variation (CV)* defined as the ratio of the standard deviation (sd) to the sample mean explains the dispersion of the data in relation to the population mean.

2. *Line length*, originally presented in [109] as an approximation of Katz's fractal dimension, is the total vertical length of the signal defined as

$$\mathrm{L}(X) = \sum_{i=2}^{N} |x_i - x_{i-1}|.$$

3. *Nonlinear energy* (*NE*), firstly established in [110], extends the concept of energy (quadratic measure) to including indefinite terms of shifted and lagged sequences, defined as

$$\mathrm{NE}(X) = \sum_{i=2}^{N-1} \left(x_i^2 - x_{i+1}x_{i-1}\right).$$

In [110], if a signal has a simple harmonic motion with the amplitude A and the oscillation frequency ω, it can be derived that NE is proportional to $A^2\omega^2$ when the sampling frequency is high. Hence, high values of NE can indicate both shifted values in a high frequency of oscillation and amplitude.

4. *Shannon entropy* (*ShEn*) [111] reflects the uncertainty in random process or quantities. It is defined as
$$\text{ShEn}(X) = -\sum_i p_i \log p_i,$$
where p_i is the probability of an occurrence of each of value in X.

5. *Approximate entropy* (*ApEn*) [112] is a measure of the regularity and fluctuation in a time series derived by comparing the similarity patterns of template vectors. The *template vector* of size m is defined as a windowed signal: $u_i = [x_i \quad x_{i+1} \quad \cdots \quad x_{i+m-1}]^T$, and we first consider the self-similarity of the template vector $u[i]$ with a tolerance r, defined by
$$C_i^m(r) = \frac{1}{N-m+1} \sum_{j=1}^{N-m+1} \Theta\left(r - \|u_i - u_j\|_\infty\right),$$
where $\Theta(x)$ is the Heaviside step function, *i.e.*, $\Theta(x)$ is one when $x \geq 0$, and zero otherwise. When X is mostly self-similar, then u_i and u_j sequences are very close and thus C_i is high. ApEn aggregates the self-similarity indices over all shifted possibilities of template vectors given template length and tolerance. ApEn is defined as
$$\text{ApEn}(X, m, r) = \frac{1}{N-m+1} \sum_{i=0}^{N-m} \log C_i^m(r) - \frac{1}{N-m} \sum_{i=0}^{N-m-1} \log C_i^{m+1}(r).$$

6. *Sample entropy* (*SampEn*) [113] is based upon a concept similar to ApEn, where SampEn compares the total number of template vectors of size m and $m+1$. SampEn differs from ApEn in which the self-similarity of *all pairs* of template vectors $u[i]$ and $u[j]$ with a tolerance r is calculated by
$$\phi^m(r) = \sum_{j=0, j\neq i}^{N-m} \sum_{i=0}^{N-m} \Theta\left(r - \|u_i - u_j\|_\infty\right).$$
If the signals are self-similar, $\phi^m(r)$ is high. SampEn is defined by
$$\text{SampEn}(X, m, r) = \log \phi^m(r) - \log \phi^{m+1}(r).$$

7. Two *Hjorth parameters*, *i.e.*, *mobility (Mob)* and *complexity (Com)*, were established to characterize the spectral properties of EEG signals in the time domain [114]. The mobility is a measure of a quantity related to the standard deviation of the signal PSD and is expressed as the ratio of the standard deviation of a signal derivative to the standard deviation of the signal:
$$\text{Mob}(X) = \frac{\text{sd}(\dot{X})}{\text{sd}(X)}$$
where $\text{sd}(X)$ is the standard deviation of X, and $\dot{X}$ denotes the first derivative of X. The complexity represents the difference between the signal and a pure sine wave and is defined by the ratio of the mobility of a signal derivative to the mobility of the signal:
$$\text{Com}(X) = \frac{\text{Mob}(\dot{X})}{\text{Mob}(X)}$$

8. *Number of zero-crossings* is an indirect measurement of the frequency characteristics of a signal. If this number is large, it means the signal contains high frequency components and more uncertainty.

9. *Number of local extrema* is the total number of local maxima and minima in a signal. It is similar to the number of zero-crossings that indirectly represents the frequency measurement of the signal.

Frequency-domain features

Frequency domain analysis is also crucial since a frequency representation of an EEG signal provides some useful information about patterns in the signal. Normally, PSD and normalized PSD (by the total power) are mostly used to extract features that represent the power partition at each frequency. This section describes features that are extracted from the PSD and the normalized PSD.

1. The *energy* is extracted from some specified frequency ranges of the PSD, usually corresponding to normal EEG activities, to determine the EEG rhythmicity in each frequency range.

2. *Intensity weighted mean frequency* (*IWMF*), also known as the mean frequency, gives the mean of the frequency distribution using the normalized PSD, and is defined as

$$\text{IWMF}\,(X) = \sum_k x_k f_k,$$

where x_k is the normalized PSD of an EEG epoch at the frequency f_k.

3. *Intensity weighted bandwidth* (IWBW) is a measure of the signal PSD width in terms of the standard deviation and is defined as

$$\text{IWBW}\,(X) = \sqrt{\sum_k x_k \left(f_k - \text{IMWF}\,(X)\right)^2},$$

where x_k is the normalized PSD. According to the seizure patterns in typical EEG signals, the PSD is sharper during seizure activities. Therefore, IWBW is smaller during those activities.

4. *Spectral entropy* (SE) is a measure of the random process uncertainty from the frequency distribution. A low SE value means the frequency distribution is intense in some frequency bands. Its calculation is similar to that for ShEn but replaces the probability distribution with the normalized PSD as follows:

$$\text{SE}(X) = -\sum_k x_k \log x_k. \tag{A.1}$$

5. *Peak frequency*, also called dominant frequency, is the frequency at which the PSD of the highest average power in its full-width-half-max band has the highest magnitude. Since the peak frequencies of normal and seizure are located differently, it can be used to differentiate the two events.

A.2 Correlation-based feature selection

CFS established in [103] is a feature selection algorithm based on the hypothesis that *a good subset of features is highly correlated with the class, but uncorrelated with others*. For a feature subset F

containing k features, an index called heuristic merit is exploited to measure feature-feature and feature-class correlations and is defined by

$$\text{Merit}_F = \frac{k\bar{r}_{fc}}{\sqrt{k + k(k-1)\bar{r}_{ff}}},$$

where $\bar{r}_{fc}$ is the mean value of the correlation of feature and class, and $\bar{r}_{ff}$ is the average of the feature-feature correlation.

To find the optimal subset F with the highest merit score, the CFS algorithm provided in [115] is applied, where the correlations are estimated by conditional entropy, and the process of this algorithm is visualized in Figure A.1. The algorithm initially assigns the subset F to be empty. New feature subsets (F_i) are constructed by adding another feature (f_i) that has not been previously selected to F. All new subsets (F_i) are then evaluated using the heuristic merit score, and the subset having the best merit score is used as the subset F in the next iteration. This process is repeated until F has m features. The final subset F contains features ranked in the descending order by the merit score.

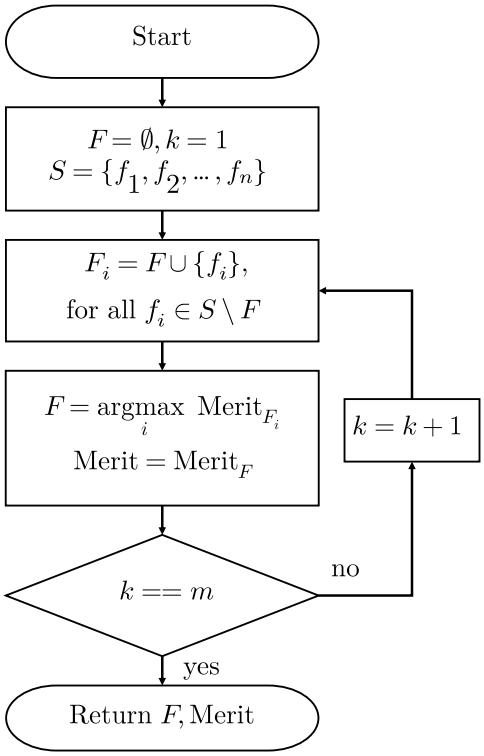


Figure A.1: Flow of CFS algorithm.

Appendix B

GRADIENT DERIVATION OF SCORENET

This section describes full details of gradient calculation of ScoreNet explained in Section 4.3. We want to find the gradients of $\mathcal{L}$ with respect to $a_1, a_2, a_3, a_4, b_1, b_2, b_3$, and b_4 since the gradients are required to update the parameters, The computational graph of ScoreNet can be visualized in Figure B. 1 where α_l and β_l are the first and last indices of the l-th candidate group. The forward computation consists of two main paths: the upper path for computing a seizure candidate c and the lower path for computing a score s. Similar to a backpropagation algorithm in neural networks, the gradients of $\mathcal{L}$ can be computed backward from an output $\hat{y}$ to an input $\mathbf{z}$.

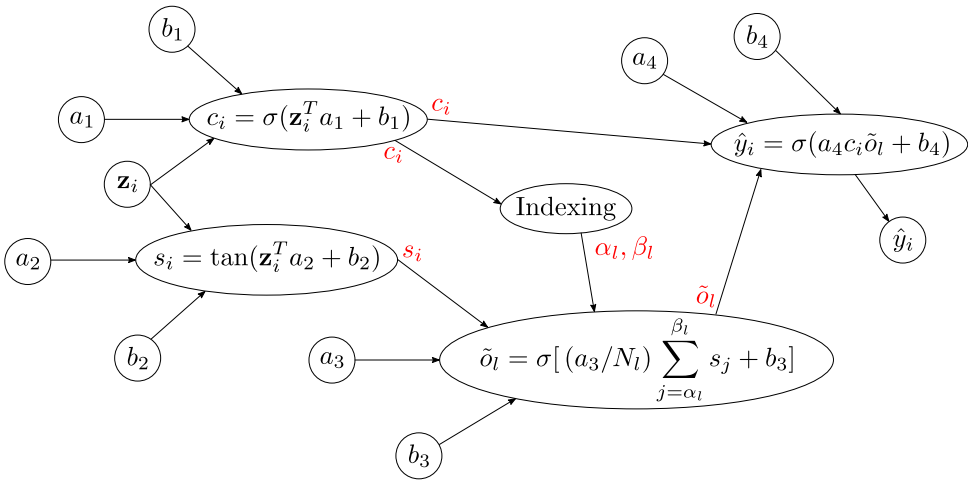


Figure B.1: Computational graph of ScoreNet. A variable in red is an output of each node used in the next node.

Consider the gradients computed at the epoch i. From (4.11) and (4.14), we can easily derive that

$$\frac{\partial \hat{y}_i}{\partial b4} = \hat{y}_i(1-\hat{y}_i), \qquad \frac{\partial \hat{y}_i}{\partial a_4} = c_i \tilde{o}_l \hat{y}_i(1-\hat{y}_i), \qquad \frac{\partial c_i}{\partial b_1} = c_i\,(1-c_i)\,, \qquad \frac{\partial c_i}{\partial a_1} = c_i\,(1-c_i)\,\mathbf{z}_i.$$

Similarly, from (4.12), the partial derivatives of s_i with respect to b_2 and a_2 are

$$\frac{\partial s_i}{\partial b_2} = 1 - s_i^2, \qquad \frac{\partial s_i}{\partial a_2} = \left(1 - s_i^2\right)\mathbf{z}_i.$$

In addition, according to (4.13), the gradients of $\tilde{o}_l$ with respect to b_3 and a_3 are

$$\frac{\partial \tilde{o}_l}{\partial b_3} = \tilde{o}_l\,(1-\tilde{o}_l) \qquad \frac{\partial \tilde{o}_l}{\partial a_3} = \frac{1}{N_l}\tilde{o}_l\,(1-\tilde{o}_l) \sum_{j \in G_l} s_j.$$

To compute the gradients of $\mathcal{L}$ with respect to the model parameters, we first split the loss $\mathcal{L}$ into a summation of local losses: $\mathcal{L}(y, \hat{y}) = \sum_{l=1}^{m} \mathcal{L}_l(y, \hat{y})$ where $\mathcal{L}_l$ is a local loss. For instance, $\mathcal{L}_l(y, \hat{y}) =$

$(1/N)\sum_{i\in G_l}[y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)]$ when the binary cross entropy is employed, and G_l contains indices of the group l. The gradients of each local loss are then calculated separately. From Figure B.1, the gradients of the local loss with respect to ScoreNet parameters in the upper path can be computed as

$$\begin{aligned}
\frac{\partial\mathcal{L}_l}{\partial b_1} &= \sum_{i\in G_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}\frac{\partial\hat{y}_i}{\partial c_i}\frac{\partial c_i}{\partial b_1} = \sum_{i\in G_l}\left[\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}a_4\tilde{o}_l\hat{y}_i(1-\hat{y}_i)c_i\,(1-c_i)\right] \\
&= a_4\tilde{o}_l\begin{bmatrix}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_{\alpha_l}} & \cdots & \frac{\partial\mathcal{L}_l}{\partial\hat{y}_{\beta_l}}\end{bmatrix}\begin{bmatrix}\hat{y}_{\alpha_l}(1-\hat{y}_{\alpha_l})c_{\alpha_l}(1-c_{\alpha_l}) \\ \vdots \\ \hat{y}_{\beta_l}(1-\hat{y}_{\beta_l})c_{\beta_l}(1-c_{\beta_l})\end{bmatrix}. \\
\frac{\partial\mathcal{L}_l}{\partial a_1} &= \sum_{i\in G_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}\frac{\partial\hat{y}_i}{\partial c_i}\frac{\partial c_i}{\partial a_1} = \sum_{i\in G_l}\left[\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}a_4\tilde{o}_i\hat{y}_i(1-\hat{y}_i)\cdot c_i\,(1-c_i)\,\mathbf{z}_i\right] \\
&= a_4\tilde{o}_l\begin{bmatrix}\mathbf{z}_{\alpha_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_{\alpha_l}} & \cdots & \mathbf{z}_{\beta_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_{\beta_l}}\end{bmatrix}\begin{bmatrix}\hat{y}_{\alpha_l}(1-\hat{y}_{\alpha_l})c_{\alpha_l}(1-c_{\alpha_l}) \\ \vdots \\ \hat{y}_{\beta_l}(1-\hat{y}_{\beta_l})c_{\beta_l}(1-c_{\beta_l})\end{bmatrix}.
\end{aligned} \tag{B.1}$$

For the lower path, according to (4.14), the derivatives of the local loss with respect to b_4 and a_4 are as follows:

$$\begin{aligned}
\frac{\partial\mathcal{L}_l}{\partial b_4} &= \sum_{i\in G_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}\frac{\partial\hat{y}_i}{\partial b_4} = \sum_{i\in G_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}\hat{y}_i\,(1-\hat{y}_i) \\
&= \begin{bmatrix}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_{\alpha_l}} & \cdots & \frac{\partial\mathcal{L}_l}{\partial\hat{y}_{\beta_l}}\end{bmatrix}\begin{bmatrix}\hat{y}_{\alpha_l}(1-\hat{y}_{\alpha_l}) \\ \vdots \\ \hat{y}_{\beta_l}(1-\hat{y}_{\beta_l})\end{bmatrix}. \\
\frac{\partial\mathcal{L}_l}{\partial a_4} &= \sum_{i\in G_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}\frac{\partial\hat{y}_i}{\partial a_4} = \sum_{i\in G_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}c_i\tilde{o}_l\hat{y}_i\,(1-\hat{y}_i) \\
&= \tilde{o}_l\begin{bmatrix}c_{\alpha_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_{\alpha_l}} & \cdots & c_{\beta_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_{\beta_l}}\end{bmatrix}\begin{bmatrix}\hat{y}_{\alpha_l}(1-\hat{y}_{\alpha_l}) \\ \vdots \\ \hat{y}_{\beta_l}(1-\hat{y}_{\beta_l})\end{bmatrix}.
\end{aligned}$$

Next, the gradient of the local loss with respect to b_3 is

$$\begin{aligned}
\frac{\partial\mathcal{L}_l}{\partial b_3} &= \sum_{i\in G_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}\frac{\partial\hat{y}_i}{\partial\tilde{o}_l}\frac{\partial\tilde{o}_l}{\partial b_3} = \sum_{i\in G_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}a_4c_i\hat{y}_i(1-\hat{y}_i)\tilde{o}_l\,(1-\tilde{o}_l) \\
&= a_4(1-\tilde{o}_l)\tilde{o}_l\sum_{i\in G_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}c_i\hat{y}_i(1-\hat{y}_i) \\
&= a_4(1-\tilde{o}_l)\frac{\partial\mathcal{L}_l}{\partial a_4}.
\end{aligned}$$

We can see that we can substitute $\frac{\partial\mathcal{L}_l}{\partial a_4}$ for $\tilde{o}_l\sum_{i\in G_l}\frac{\partial\mathcal{L}_l}{\partial\hat{y}_i}c_i\hat{y}_i(1-\hat{y}_i)$ so that these terms are not repeatedly computed. Consequently, the gradients of the local loss with respect to the other ScoreNet parameters

associated with the lower path are demonstrated below.

$$\begin{aligned}
\frac{\partial \mathcal{L}_l}{\partial a_3} &= \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial \tilde{o}_l} \frac{\partial \tilde{o}_l}{\partial a_3} = \sum_{i \in G_l} \left[\frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} a_4 c_i \hat{y}_i (1 - \hat{y}_i) \tilde{o}_l \left(1 - \tilde{o}_l\right) \frac{1}{N_l} \sum_{j \in G_l} s_j \right] \\
&= \frac{1}{N_l} \left[a_4 (1 - \tilde{o}_l) \tilde{o}_l \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} c_i \hat{y}_i (1 - \hat{y}_i) \right] \left(\sum_{j \in G_l} s_j \right) \\
&= \frac{1}{N_l} \frac{\partial \mathcal{L}_l}{\partial b_3} \sum_{j \in G_l} s_j. \qquad \text{(B.2)} \\
\frac{\partial \mathcal{L}_l}{\partial b_2} &= \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial \tilde{o}_l} \frac{\partial \tilde{o}_l}{\partial \sum s_j / N_l} \frac{\partial \sum s_j / N_l}{\partial b_2} \\
&= \sum_{i \in G_l} \left[\frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} a_4 c_i \hat{y}_i (1 - \hat{y}_i) \tilde{o}_l \left(1 - \tilde{o}_l\right) \frac{a_3}{N_l} \sum_{j \in G_l} (1 - s_j)^2 \right] \\
&= \frac{a_3}{N_l} \left[a_4 (1 - \tilde{o}_l) \tilde{o}_l \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} c_i (\hat{y}_i - y_i) \right] \left(\sum_{j \in G_l} (1 - s_j^2) \right) \\
&= \frac{a_3}{N_l} \frac{\partial \mathcal{L}_l}{\partial b_3} \sum_{j \in G_l} (1 - s_j^2). \qquad \text{(B.3)} \\
\frac{\partial \mathcal{L}_l}{\partial a_2} &= \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial \tilde{o}_l} \frac{\partial \tilde{o}_l}{\partial \sum s_j / N_l} \frac{\partial \sum s_j / N_l}{\partial a_2} \\
&= \sum_{i \in G_l} \left[\frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} a_4 c_i \hat{y}_i (1 - \hat{y}_i) \tilde{o}_l \left(1 - \tilde{o}_l\right) \frac{a_3}{N_l} \sum_{j \in G_l} \left[(1 - s_j^2) \mathbf{z}_j \right] \right] \\
&= \frac{a_3}{N_l} \left(a_4 (1 - \tilde{o}_l) \tilde{o}_l \sum_{i \in G_l} \frac{\partial \mathcal{L}_l}{\partial \hat{y}_i} c_i (\hat{y}_i - y_i) \right) \left(\sum_{j \in G_l} (1 - s_j^2) \mathbf{z_j} \right) \\
&= \frac{a_3}{N_l} \frac{\partial \mathcal{L}_l}{\partial b_3} \left[\mathbf{z}_{\alpha_l} \quad \cdots \quad \mathbf{z}_{\beta_l} \right] \begin{bmatrix} \left(1 - s_{\alpha_l}^2\right) \\ \vdots \\ \left(1 - s_{\beta_l}^2\right) \end{bmatrix}. \qquad \text{(B.4)}
\end{aligned}$$

These calculations are similar to backpropagation in neural network, and we can see that some terms are repeatedly used for calculation. For instance, $\frac{\partial \mathcal{L}_l}{\partial b_3}$ is used in (B.2), (B.3), and (B.4). Hence, we can compute these repeated terms once and store them to reduce time for the calculations. Finally, the gradients of $\mathcal{L}$ are obtained by combining the gradients of $\mathcal{L}_l$, *i.e.*, $\nabla \mathcal{L} = \sum_{l=1}^{m} \nabla \mathcal{L}_l$.

Appendix C

SUPPLEMENTARY RESULTS

This chapter presents supplementary results used for supporting the main experiments in Chapter 6. First, apart from the CHB-MIT Scalp EEG database, the significance of each feature is also explored using an EEG record chosen from King Chulalongkorn Memorial Hospital. The EEG record for this study has passed the Institutional Review Board of Faculty of Medicine, Chulalongkorn university. An experiment of finding the most useful parameters m and r in approximate entropy and sample entropy is conducted in Appendix C.2. Appendix C.3 discusses effects of parameters of the proposed counting-based method on seizure detection performances. Moreover, we explain optimization setups including initial point selection of ScoreNet in Appendix C.4. Finally, a correlation between detection accuracy and an imbalance ratio of the training set is provided in Appendix C.5 to demonstrate a key factor to the detection performance.

C.1 Feature significance using private record

This experiment is conducted to support the experiment of feature analysis described in Section 6.1 using a different EEG data set to ensure that attributes which are related to changes in amplitude, frequency, and rhythmicity of EEGs achieve some significant levels.

Experiment

All details of the experimental setup were the same as in Section 6.1 except the data set. In this experiment, the private record of 25 minutes from King Chulalongkorn Memorial Hospital was used to analyze each individual feature. This record was collected with a sampling rate of 200 Hz measured with the international 10-20 system. All seizure activities in this record were chosen to be generalized 3-Hz spike-slow wave seizures in order to easily analyze and interpret the significance of each feature.

According to the chosen record, the number of normal epochs is $1{,}066$, the among of seizure epoch is 276, and the total number of epoch is $1{,}342$. Therefore, the error of the common situation is $\text{err}_0 = 0.1699$. The features with improvement rate more than 40%, emphasized as boldface, were regarded as significant. Table C.1 shows the Bayesian errors and improvement rates of all time-domain and frequency-domain features on left and right half of the brain. Apparently in time domain, *variance, max, min, energy, nonlinear energy, Shannon entropy, approximate entropy, sample entropy, and complexity* achieved typically high improvement rates of more than 40% for both left and right sides, where variance yielded the best results of 67.30% and 55.00% improvement rates from the left and right hemispheres, respectively. In frequency domain, IWBW and peak amplitude achieved similar improvement rates around 40% on the left side, whereas their improvement rates were relatively lower on the right side. This is because the spectral density of a seizure epoch is intense in some frequency range because of the rhythmic patterns of spike-slow wave seizures.

Figure C.1a demonstrates the improvement rates of time-frequency-domain features in each decomposition level. We found that amplitude-related features such as variance, energy, and line length from D3, D4, and D5 gained improvement rates of over 60% on both sides, and these features computed from the fifth approximation level on the left half brain obtained high improvement rates. On the other hand, mean, skewness, and kurtosis from the decomposition levels obtained relatively low improvement rates of less than 40%. It was also found that all features extracted from D1 achieved low improvement

Table C.1: Bayes error (err_b) and improvement rate of time-domain and frequency-domain features using private record.

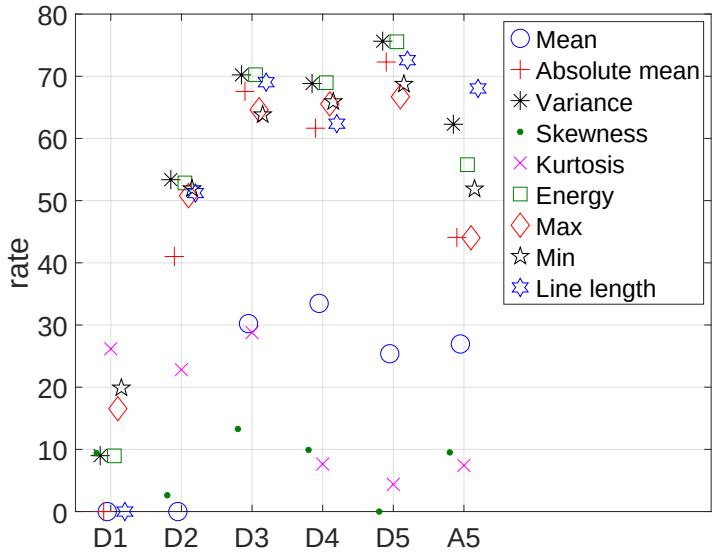
(a)) **Time-domain** features.

Feature	Left side		Right side	
	err_b	**rate**	err_b	**rate**
Mean	0.1253	26.22	0.1284	24.44
Variance	0.0556	**67.30**	0.0764	**55.00**
CV	0.1606	5.47	0.1589	6.50
Skewness	0.1589	6.48	0.1452	14.54
Kurtosis	0.1588	6.54	0.1574	7.35
Max	0.0649	**61.80**	0.0837	**50.73**
Min	0.0854	**49.76**	0.0947	**44.28**
Energy	0.0624	**63.27**	0.1075	36.74
NE	0.0640	**62.34**	0.0765	**54.95**
Line Length	0.1034	39.14	0.1237	27.20
H_{sh}	0.0844	**50.31**	0.0978	**42.43**
ApEN	0.0949	**44.14**	0.1030	39.37
SampEN	0.0889	**47.68**	0.0835	**50.88**
Local extrema	0.1370	19.34	0.1257	26.01
Zero-crossing	0.1699	0.00	0.1699	0.00
Mobility	0.1370	19.36	0.1221	28.12
Complexity	0.0962	**43.35**	0.0878	**48.33**

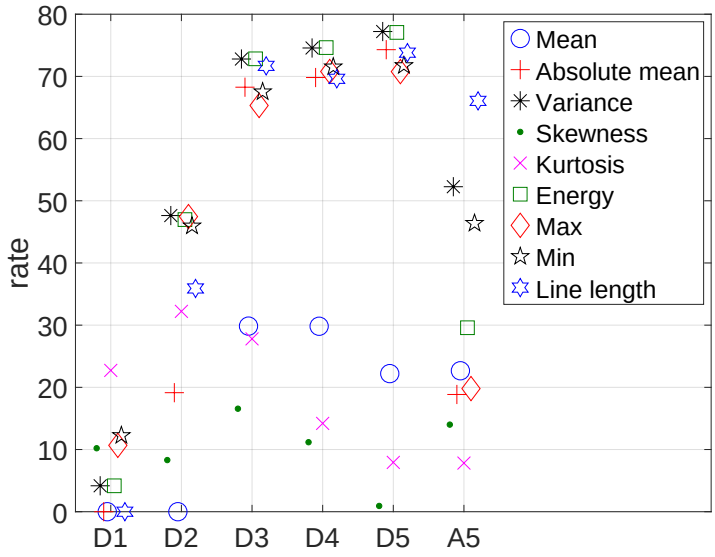
(b)) **Frequency-domain** features.

Feature	Left side		Right side	
	err_b	**rate**	err_b	**rate**
IWMF	0.1699	0.00	0.1699	0.00
IWBW	0.1111	34.63	0.0993	**41.58**
SE	0.1699	0.00	0.1699	0.00
Peak amplitude	0.0924	**45.62**	0.1435	15.53
Peak frequency	0.1415	16.70	0.1241	26.94

rates, whereas the improvement rates of those features from A5 varied depending on sides of the brain. According to the decomposition technique, we found that decomposition levels D3, D4, and D5 correspond to rhythms of normal EEG, whereas the fifth approximation level contains information of a low frequency range. In this record, magnitudes of the signal in ictal patterns on the right side are more dominant than those on the left side; thus, amplitude-related features from the right side are more significant. In addition, the decomposition level D1 which corresponds to frequency band of higher than 50 Hz does not provide any information related to changes in EEGs. As a result, features that reflect changes in amplitude, frequency, and rhythmicity of EEGs are capable of being used for the seizure detection. Additionally, time-frequency-domain features are most significant in specific sub-bands as



(a) *left* hemisphere.



(b) *right* hemisphere.

Figure C.1: Improvement rates of time-frequency domain features calculated on private record. Labels `D1`, `D2`, `D3`, `D4`, `D5`, and `A5` represent sub-bands from which the features are extracted.

more information can be obtained from the sub-band.

C.2 ApEn and SampEn parameter selection

Approximate entropy (ApEn) and sample entropy (SampEn) are measurements of uncertainty and predictability over time-series data, as described in Appendix A.1. ApEn and SampEn calculated

from a time series data which consists of similar, predictable, and periodic pattern has a relatively small value, whereas a more uncertain, unpredictable time series gives higher ApEn and SampEn. However, an appropriate choice of a template length m and a tolerance r largely depends on an application. In the application of the epileptic seizure detection, several values have been suggested without supports from experiments. Therefore, this experiments aims to find the template length and the tolerance so that ApEn and SampEn can powerfully distinguish between normal and ictal patterns in EEGs.

According to predictability of EEGs during seizures, ApEn and SampEn of seizure EEG epochs are normally lower than those of normal EEG epochs, and the difference of ApEn and SampEn from the normal and abnormal EEGs largely depend on m and r. In order to obtain ApEn and SampEn that can potentially differentiate between the EEGs of normal and ictal patterns, m and r are selected to maximize the differences of the features between these classes:

$$(m_{\max}, r_{\max}) = \underset{m,r}{\operatorname{argmax}} \ \left|\text{SampEn}_{\text{norm}}(m,r) - \text{SampEn}_{\text{epi}}(m,r)\right|,$$

and

$$(m_{\max}, r_{\max}) = \underset{m,r}{\operatorname{argmax}} \ \left|\text{ApEn}_{\text{norm}}(m,r) - \text{ApEn}_{\text{epi}}(m,r)\right|,$$

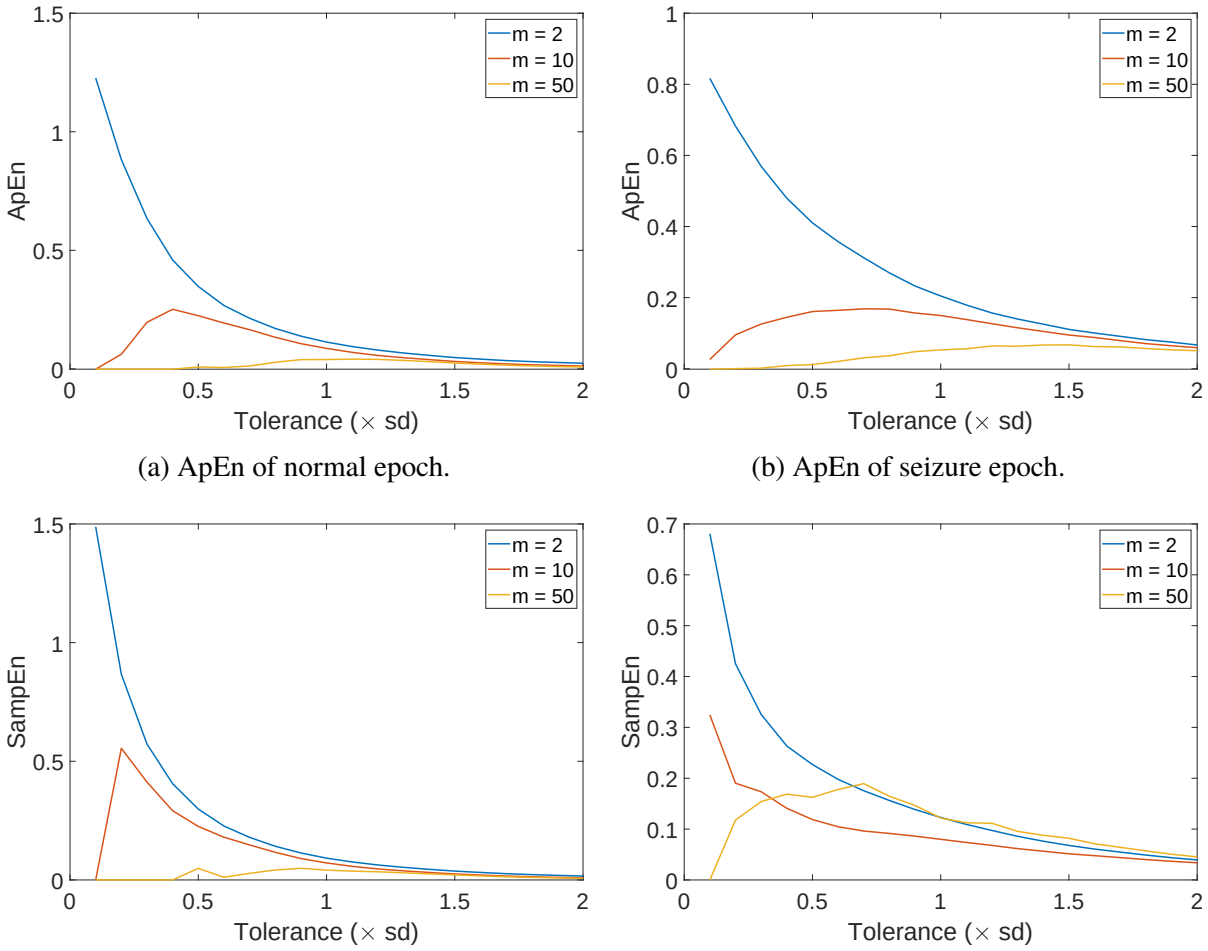
where $\text{SampEn}_{\text{norm}}(m,r)$, $\text{ApEn}_{\text{norm}}(m,r)$, $\text{SampEn}_{\text{epi}}(m,r)$, and $\text{ApEn}_{\text{epi}}(m,r)$ are SampEn and ApEn from normal and abnormal epochs, respectively.

Experiment

In this experiment, a single EEG signal used in this experiment was selected from the CU database. Since the record contained only generalized 3-Hz spike-and-slow wave seizures, using only one EEG channel is sufficient to analyze ApEn and SampEn. Therefore, ApEn and SampEn were extracted from the first channel (*FP1-F7*). Thirty epochs of four seconds from each of the normal and abnormal classes were randomly chosen. Each epoch was used to extract ApEn and SampEn using various pairs of m and r. A template length m and a tolerance r were varied independently: the template length was varied from 2 to 80 sample points, and the tolerance was adjusted from 0.1sd to 2sd by 0.1sd where sd stands for the standard deviation of the EEG epoch. A combination of a feature from these two classes was then constructed to find the optimal m and r for each combination; hence, the total number of these combinations was 900. However, it is possible that SampEn is indefinite when $\phi^m(r) = \phi^{m+1}(r) = 0$: there is no occurrence of matching in the signal, and when $\phi^m \neq 0, \phi^{m+1}(r) = 0$: there is no matching with the template vector of length $m+1$. In this case, we neglected the invalidity case of SampEn; thus, the number of combinations of $\text{SampEn}_{\text{norm}}$ and $\text{SampEn}_{\text{epi}}$ was less than 900. Finally, a histogram of the optimal m and r of each combination was constructed to find the most suitable m and r that provided the highest number of the largest difference of each feature between these two classes.

Figure C.2 showing effects of a template length m and a tolerance r on ApEn and SampEn verifies that ApEn and SampEn from a normal epoch are typically higher than those of a seizure epoch. Decreases in both ApEn and SampEn occurred when the tolerance exceeded a certain value. It means that the numbers of vector pairs of length m and $m+1$ are dissimilar when the tolerance is small. On the other hand, when the tolerance is higher than the certain value, the difference between $\phi^m(r)$ and $\phi^{m+1}(r)$ are both higher; hence, ApEn and SampEn are smaller. In addition, ApEn and SampEn were also reduced when the template length increased. This indicates that a longer template length implies a more difficulty for matching a pattern of a template vector in the input signal. As a result, $\phi^m(r)$ and $\phi^{m+1}(r)$ are similarly smaller, and ,thus, ApEn and SampEn are lower.

Figure C.3 demonstrates the number of the maximum differences of ApEn (Figure C.3a) and SampEn (Figure C.3b) between seizure and normal signals. We can see that almost all maximum differences were discovered when the tolerance r was in a range of 0.1sd and 0.2sd, and the template length m was low. Specifically, in the case of ApEn, the maximum differences mostly occurred at $m = 2$ and



(a) ApEn of normal epoch.

(b) ApEn of seizure epoch.

(c) SampEn of normal epoch.

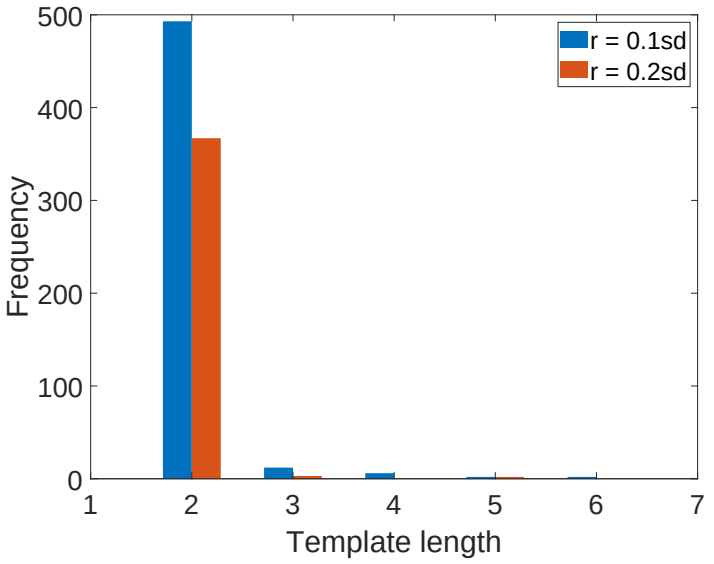
(d) SampEn of seizure epoch.

Figure C.2: ApEn and SampEn when r is varied for a fixed m.

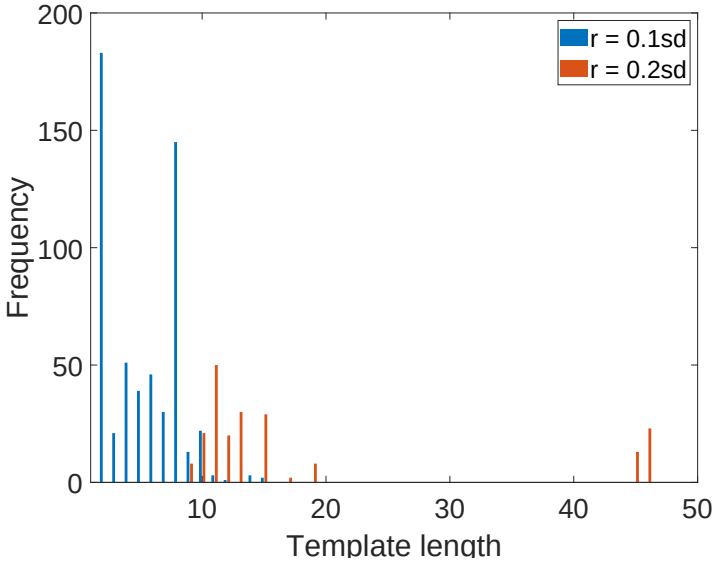
$r = 0.1$sd, whereas another large portion appeared at $r = 0.2$sd. This means that the maximum difference of ApEn from normal and seizure epochs is considerably reduced when r and m are higher, as shown in Figures C.2a and C.2b. On the other hand, the maximum differences in the SampEn case were distributed along the axis $r = 0.1$sd, and the highest frequency was located at $m = 2$. As a result, from the chosen data, we can conclude that $m = 2$ and $r = 0.1$sd are the most suitable parameters of ApEn and SampEn that help differentiate between normal and seizure epochs.

C.3 Parameter selection of counting-based method

Explained in Section 4.3, the counting-based method is a criterion-based post-processing technique for indicating the starting and ending time points of an epileptic seizure. There are two parameters, *i.e.*, w and p, that need to be specified. In this section, we perform an experiment to analyze effects of these parameters on sensitivity, specificity, and F_1. Furthermore, the parameters w and p achieving the highest F_1 are then used in Section 6.2. Since this work was already published [17], some details are taken from the study.



(a) ApEn.



(b) SampEn.

Figure C.3: Histograms of m and r obtained from maximizing the differences of ApEn and SampEn from normal and seizure epochs.

Experiment

In this experiment, we applied the counting-based method to outcomes (normal/seizure) of the proposed CNN model shown in Figure 4.2 from all test records. For a specific w and p, sensitivity, specificity, and F_1 were collected from each test record as performance metrics. These measurements were averaged over records of each patient and then averaged again to obtain the overall metrics. The parameters w and p were varied from 0 to 300 and from 1 to 30, respectively. Subsequently, the overall sensitivity and specificity obtained from each case were used to display a receiver operating characteristic (ROC) curve and analyze the effects of these parameters on the counting-based method.

ROC curves when varying w with different values of p are demonstrated in Figure C.4. We found that the sensitivity was significantly affected by w and p, whereas the specificity only changed less than 10%. When w was higher, the sensitivity rapidly increased at the beginning and converged to a certain value after w exceeded a certain point; in contrast, the sensitivity decreased as opposed to p. This implies that the method is more capable of detecting seizures when increasing w and decreasing p. When a required gap between detected epochs w is large, and only a few number of repeated detected epochs p is needed, the criterion of allowing seizure candidates to be regarded as seizures (4.7) can be easily satisfied. However, the specificity still decreased when w exceeded the certain point. As explained in Section 4.3, when w is higher, more epochs near a detected epoch are allowed to be seizure candidates, leading to a more number of false positives. Some of these candidates that are not parts of a seizure are eventually declared as seizure epochs because the criterion (4.7) is satisfied. As a result, this implies that a high w does provide terrible seizure detection outcomes, and, thus, appropriate values of w and p are required.

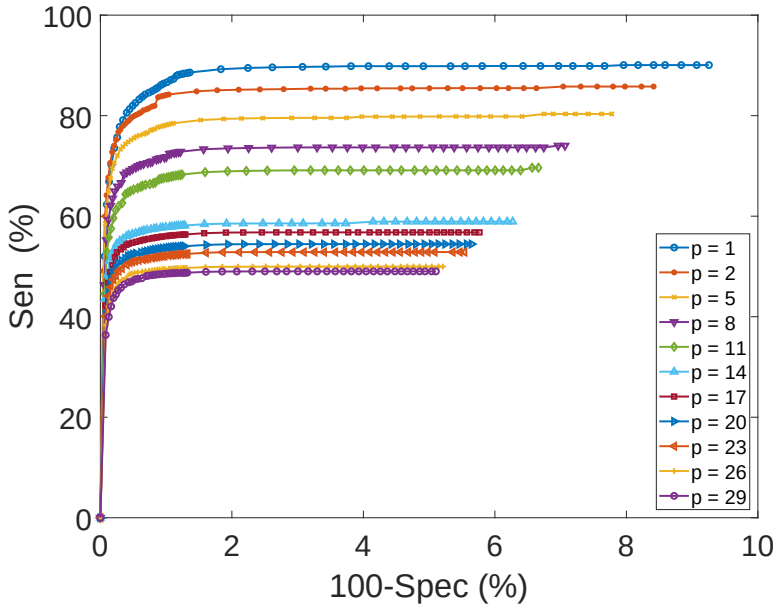


Figure C.4: ROC curve when p is fixed and w is varied.

In addition, averaged F_1 from all test cases for a specific p is plotted in Figure C.5. It is evident that an increase of p typically reduced F_1. For any p, when w increased, F_1 increased at the beginning, and was maximum at a certain point. When w was higher than the certain value, F_1 was deducted. As discussed above, a small value of p allows the requirement (4.7) to be easily satisfied, yielding high F_1. On the other hand, F_1 decreased as w was larger since more excessive false negatives appeared. Finally, as illustrated in Figure C.5, $w = 6$ and $p = 2$ providing the highest F_1 of 64.40% are the most suitable parameters the counting-based method.

Finally, Figure C.6 compares averaged F_1 scores obtained from only CNN and CNN with the counting-based method. Note that the averaged F_1 scores from CNN across patients are largely different, ranging from 2.56% to 57.66%. This means that the detection performances are highly relevant to the demographics of patients. According to various types of seizures observed from different patients, it might be extremely hard (or even impossible) to tune hyperparameters of CNN that are optimal for all patients. These preliminary results lead to our decision on the patient-specific scheme as explained in Section 1.2.

It is remarkable that, except the case `chb04`, the averaged F_1 score of each case is higher after using the counting-based method. This means that the counting-based is capable of solving problems of isolated false positives and false negatives. Nevertheless, for `chb04`, F_1 obtained from the counting-

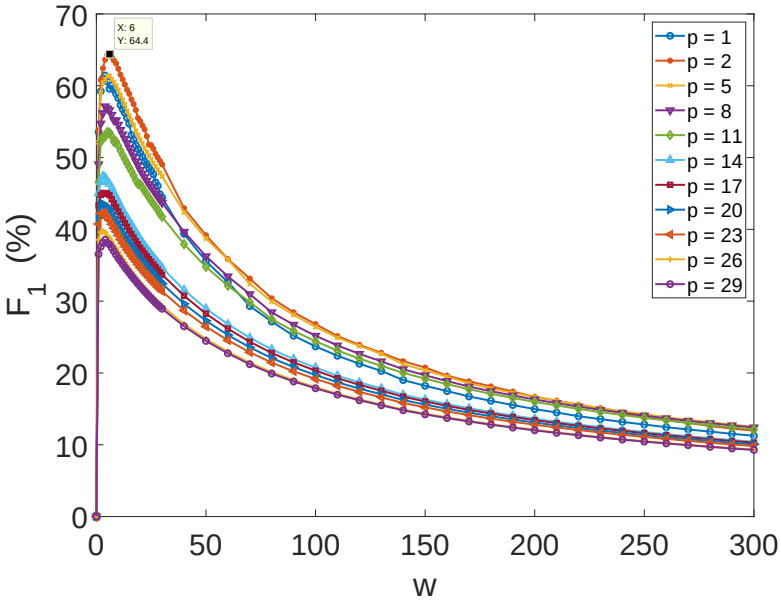


Figure C.5: F_1 score when p is fixed and w is varied.

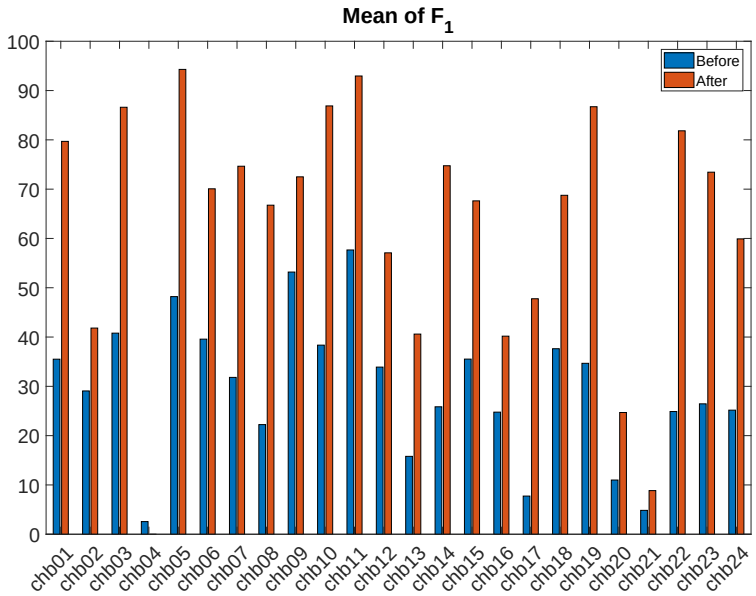


Figure C.6: Averaged F_1 of each subject before and after using the counting-based method.

based method vanished because CNN could detect only one seizure epoch. As a result, the counting-based method ignored this epoch because of an insufficient number of near detected epochs.

C.4 Initialization of ScoreNet parameters

This section explains how we setup ScoreNet used in this dissertation. We firstly set the size of each ScoreNet parameter corresponding to the optimal w in order to compare the performance with the best counting-based method; the size of vectors a_1 and a_2 was 13, and the other parameters are scalars. Three initial points were then selected due to the prior classification results based on the training data. From (B.4) and (B.1), as z_i is an output of a classification problem of highly imbalanced data, z is extremely sparse; $\mathbf{z}_i$ is usually a zero vector, and the gradients become vanished. Thus, the parameters a_1 and a_2 are slightly updated, and hardly converge to a good local minimum. In this case, we mainly initialized a_1 and a_2 first based on three situations, and the initialization of the other parameters was empirically obtained to avoid a convergence to poor optimal points. Initial points of a_1 and a_2 were based on the parameters of the counting based method, *i.e.*, the initial points were multiples of $\mathbf{1}$. When there are many false negatives from the prior classification step, magnitudes of a_1 and a_2 should be high to boost a seizure candidate c_i and we easily assign $s_i = 1$. On the other hand, the magnitudes of a_1 and a_2 should be small to suppress the effect of false positives when the number of isolated false positive is dominant. In addition, a_1 and a_2 should have the intermediate magnitudes relatively compared to these two cases if the numbers of false positives and negatives are slightly different. The initial values according to the circumstances are eventually listed in Table C.2.

Table C.2: List of initial parameters of ScoreNet. Vector of ones is represented by $\mathbf{1}$.

Cases	**Parameters**							
	a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4
High false positive	$\mathbf{1}$	$\mathbf{1}$	3	3	-1	-2	0	-1
High false negative	$9 \cdot \mathbf{1}$	$9 \cdot \mathbf{1}$	4	4	-4	-8	1	-1
Balance	$6 \cdot \mathbf{1}$	$6 \cdot \mathbf{1}$	3	3	-3	-9	0	-1

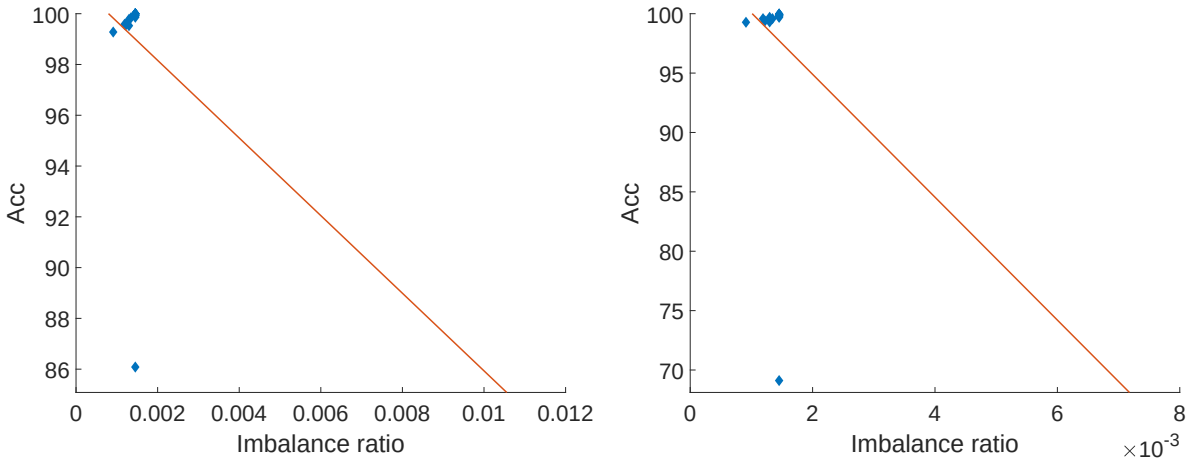
C.5 Key factor to detection performance

Classification performances are naturally affected by a portion of training class sizes, especially in the case of imbalanced data. Predictions from a classification method tend to be biased toward the majority class, *i.e.*, a normal class in this case. Many previous studies chose fewer normal samples to reduce the difference between class sizes without any statistical evidence. In this section, a correlation coefficient between a performance metric and an *imbalance ratio* is exploited to demonstrate a key factor to the classification performance. The imbalance ratio is a ratio of the seizure class size to the normal class size where the ratio is zero when there are no abnormal samples, and one when the data are balanced. When the ratio increases, a number of true positives is expected to be higher. Accuracy is used for the analysis because it considers both true positives and true negatives, so it suitably reflects the effect of the data imbalance. Unlike F_1 scores that can be calculated from only limited records containing seizures, leading to insignificant and biased results, the accuracy is obtained from every test case. Thus, the correlation coefficient between the accuracy and the imbalance ratio is supposed to be significantly positive.

Experiment

In this experiment, the CHB-MIT Scalp EEG database was used to determine effects of the imbalance ratio on the accuracy. Epoch-based seizure detectors, *i.e.*, CNN, logistic regression, SVM, decision tree, and random forest were exploited with the same setup explained in Section 6.2. Here, the accuracy from a test case was collected with the imbalance ratio of the training set. The correlation coefficient between the accuracy and the ratio was subsequently calculated for each patient.

Table C.3 shows correlation coefficients between the accuracy and the imbalance ratio with the corresponding p-values. It is evident from the p-values ($\alpha = 0.05$) that the correlation coefficients were all significantly positive except the cases of applying the SVM and logistic regression to patient `chb16`. On the other hand, for the negative correlation cases, the coefficients were slightly negative and the p-values were high, meaning that the correlations were weak. We observed that low accuracy was obtained from a single record (`chb16_01`) as demonstrated in Figure C.7. Several isolated false positives were predicted, resulting in remarkably low accuracy and insignificantly negative correlation. As a result, the detection accuracy generally tends to be higher when more abnormal samples are added. Therefore, the imbalance ratio of the training set is a key factor to the detection performance.



(a) Logistic regression.

(b) SVM.

Figure C.7: Illustration of correlation between accuracy and imbalance ratio for logistic regression and SVM of the case `chb16` where red lines indicate trends using linear regressions.

Table C.3: Correlation coefficients between accuracy and imbalance ratio from each patient. Numbers in parentheses are the corresponding p-values, and red numbers indicate insignificant correlations ($p \geq 0.05$).

Patient	CNN	Logistic regression	SVM	Decision tree	Random forest
chb01	0.8777	0.7175	0.1533	0.8684	0.8580
	(2×10^{-14})	(9×10^{-8})	(0.3323)	(9×10^{-14})	(4×10^{-13})
chb02	0.9235	0.8735	0.9404	0.9883	0.9609
	(1×10^{-15})	(4×10^{-12})	(2×10^{-17})	(2×10^{-29})	(2×10^{-20})
chb03	0.8379	0.4903	0.2855	0.8614	0.8881
	(5×10^{-11})	(0.0018)	(0.0823)	(4×10^{-12})	(1×10^{-13})
chb04	0.9959	0.4323	0.7276	0.9832	0.9947
	(3×10^{-43})	(0.0043)	(5×10^{-8})	(4×10^{-31})	(4×10^{-41})
chb05	0.8403	0.9475	0.9507	0.9472	0.9164
	(2×10^{-11})	(6×10^{-20})	(2×10^{-20})	(7×10^{-20})	(3×10^{-16})
chb06	0.7897	0.1783	0.1781	0.8873	0.9369
	(0.0001)	(0.4790)	(0.4794)	(9×10^{-7})	(1×10^{-8})
chb07	0.5636	0.5573	0.6072	0.6444	0.5542
	(0.0120)	(0.0132)	(0.0058)	(0.0029)	(0.0138)
chb08	0.9713	0.8916	0.8133	0.9108	0.9727
	(1×10^{-12})	(1×10^{-7})	(1×10^{-5})	(2×10^{-8})	(7×10^{-13})
chb09	0.4885	0.4478	0.4189	0.8431	0.9823
	(0.0338)	(0.0545)	(0.0742)	(6×10^{-6})	(8×10^{-14})
chb10	0.4973	0.7138	0.4408	0.4683	0.7960
	(0.0114)	(0.0001)	(0.0274)	(0.0182)	(2×10^{-6})
chb11	0.9993	0.9379	0.9942	0.9996	0.9997
	(3×10^{-48})	(1×10^{-16})	(2×10^{-33})	(2×10^{-52})	(3×10^{-55})
chb12	0.7815	0.8538	0.8595	0.8912	0.8805
	(7×10^{-6})	(1×10^{-7})	(7×10^{-8})	(5×10^{-9})	(1×10^{-8})
chb13	0.9897	0.9195	0.8326	0.8741	0.9998
	(9×10^{-28})	(4×10^{-14})	(2×10^{-9})	(3×10^{-11})	(2×10^{-55})
chb14	0.9211	0.8652	0.8038	0.9695	0.9639
	(3×10^{-11})	(1×10^{-8})	(8×10^{-7})	(4×10^{-16})	(3×10^{-15})
chb15	0.5120	0.5471	0.5345	0.5213	0.5779
	(0.0007)	(0.0003)	(0.0004)	(0.0006)	(0.0001)
chb16	0.9816	−0.0688	−0.1047	0.9919	0.9981
	(1×10^{-13})	(0.7797)	(0.6696)	(1×10^{-16})	(5×10^{-22})
chb17	0.9636	0.8547	0.8412	0.9879	0.9998
	(2×10^{-12})	(8×10^{-7})	(2×10^{-6})	(7×10^{-17})	(2×10^{-33})
chb18	0.4052	0.4349	0.5905	0.6308	0.3734
	(0.0142)	(0.0080)	(0.0001)	(4×10^{-5})	(0.0249)
chb19	0.9983	0.8942	0.8591	0.9319	0.9078
	(6×10^{-36})	(3×10^{-11})	(1×10^{-9})	(7×10^{-14})	(4×10^{-12})
chb20	0.8810	0.8128	0.7631	0.9326	0.9340
	(3×10^{-10})	(8×10^{-8})	(1×10^{-6})	(2×10^{-13})	(1×10^{-13})
chb21	0.8680	0.9119	0.7093	0.9889	0.9999
	(6×10^{-11})	(2×10^{-13})	(4×10^{-6})	(3×10^{-27})	(6×10^{-59})
chb22	0.9574	0.9422	0.9375	0.9676	0.9931
	(4×10^{-17})	(3×10^{-15})	(8×10^{-15})	(7×10^{-19})	(1×10^{-28})
chb23	0.8949	0.8798	0.8545	0.9329	0.9210
	(0.0011)	(0.0018)	(0.0033)	(0.0002)	(0.0004)
chb24	0.7961	0.3341	0.3595	0.3785	0.4239
	(9×10^{-6})	(0.1286)	(0.1003)	(0.0824)	(0.0493)

Appendix D

LIST OF PUBLICATIONS

1. P. Boonyakitanont, A. Lek-uthai, K. Chomtho, and J. Songsiri, "A comparison of deep neural networks for seizure detection in EEG signals," *bioRxiv preprint*, p. 702654, 2019, https://www.biorxiv.org/content/10.1101/702654v1

2. ——, "A review of feature extraction and performance evaluation in epileptic seizure detection using EEG," *Biomedical Signal Processing and Control*, vol. 57, p. 101702, 2020

3. P. Boonyakitanont, A. Lek-uthai, and J. Songsiri, "Automatic epileptic seizure onset-offset detection based on CNN in scalp EEG," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1225–1229

4. P. Boonyakitanont, A. Lek-uthai, and J. Songsiri, "ScoreNet: A neural network-based post-processing model for identifying epileptic seizure onset and offset in EEGs," *bioRxiv preprint*, 2020, https://www.biorxiv.org/content/10.1101/2020.12.21.423728v2

Biography

Poomipat Boonyakitanont received his Bachelor's of Engineering (2nd Class Honours) in Electrical Engineering from Chulalongkorn University, Thailand, in 2015. He was an intern as a maintenance engineer at Federal Electric Co., Ltd., and he was employed in Dynoelectric Co., Ltd. as a sourcing engineer after graduated. In 2017, he began his Ph.D. in Biomedical Engineering, Department of Electrical Engineering, Chulalongkorn University. He was financially supported by a grant titled *the 100th Anniversary Chulalongkorn University for Doctoral Scholarship* since his graduation program. In 2020, he received *the 90th Anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund)*. He was awarded *the 2020 Fulbright Junior Research Scholarship Program* to conduct research at University of Maryland, Baltimore County, U.S. His current research of interest includes various applications of machine learning and deep learning such as biomedical data analysis, signal processing, and object detection.