Titre: Title:	Efficient Deep Learning Algorithms for Robust Prediction of Epileptic Seizures
Auteur: Author:	Yang Zhang
Date:	2024
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Zhang, Y. (2024). Efficient Deep Learning Algorithms for Robust Prediction of Epileptic Seizures [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/62505/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/62505/
Directeurs de recherche: Advisors:	François Leduc-Primeau, Yvon Savaria, & Mohamad Sawan
Programme: Program:	Génie électrique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Efficient Deep Learning Algorithms for Robust Prediction of Epileptic Seizures

YANG ZHANG
Département de génie électrique

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie électrique

Novembre 2024

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

Efficient Deep Learning Algorithms
for Robust Prediction of Epileptic Seizures

présentée par **Yang ZHANG**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Frédéric LESAGE, président
François LEDUC-PRIMEAU, membre et directeur de recherche
Mohamad SAWAN, membre et codirecteur de recherche
Yvon SAVARIA, membre et codirecteur de recherche
Farida CHERIET, membre
Yong LIAN, membre externe

DEDICATION

To my beloved grandma,
Per aspera ad astra...

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to Prof. Mohamad Sawan, who gave me the chance to pursue my Ph.D. program. I am also extremely grateful and fortunate to have been under the supervision of Prof. François Leduc-Primeau and Prof. Yvon Savaria for their invaluable guidance and mentorship. All your steadfast support and careful guidance were crucial to my success. I would not have reached the end of this journey without all your patience, motivation, and immense knowledge.

My sincere thanks also extend to the jury member, Prof. Frédéric Lesage, Prof. Farida Cheriet, and Prof. Yong Lian, for generously dedicating their time despite their busy schedules.

Time has truly flown by. I fondly recall the joyful moments with all my colleagues and friends spent at the Polystim Neurotech Laboratory and the Microelectronics and Microsystems Research Group. I would like to extend my heartfelt thanks to everyone for their invaluable help and guidance.

Besides, I am deeply grateful to IVADO, CMC Microsystems, and the Digital Research Alliance of Canada for their technical support and for providing essential computing resources.

Last but not least, I am profoundly grateful to my family and beloved for their unconditional love and unwavering support.

RÉSUMÉ

La prédiction des crises est importante pour les personnes atteintes d'épilepsie pharmacorésistante, qui peuvent ne pas bien répondre aux médicaments antiépileptiques (AEDs). Pour ces patients, elle améliore considérablement la sécurité en permettant aux individus de prendre des mesures de précaution ou des interventions alternatives contre les blessures potentielles associées à des crises inattendues. De plus, la prédiction des crises permet de minimiser le recours à des médicaments qui peuvent avoir de graves effets secondaires. De nombreux AEDs s'accompagnent de divers effets secondaires qui peuvent affecter la qualité de vie des patients. En permettant un meilleur timing et en réduisant la dose de médicament grâce à une prédiction précise, les patients peuvent ressentir moins d'effets secondaires tout en maintenant un contrôle efficace des crises. Les dispositifs portables ou implantables sont donc essentiels dans la prise en charge de l'épilepsie, car ils permettent une surveillance continue et en temps réel des signaux d'électroencéphalogramme (EEG). Cette capacité permet de détecter les schémas précurseurs des crises, ce qui permet de déclencher des alertes et des interventions précoces susceptibles d'améliorer l'autonomie et la qualité de vie des patients.

Cette thèse porte sur un moteur de prédiction de crises d'épilepsie robuste adapté aux dispositifs médicaux portables afin d'explorer des approches pratiques pour aider les patients épileptiques à obtenir des informations de prédiction de crises à temps. L'objectif principal de ce travail est le développement de méthodes performantes, économes en énergie et interprétables pour la prédiction des crises dans des dispositifs portables à ressources limitées qui peuvent potentiellement convenir à une utilisation quotidienne. Le principal défi de ce travail réside dans le compromis entre performances et consommation d'énergie, mettant en évidence un dilemme permanent dans les dispositifs médicaux portables à ressources limitées. De plus, l'objectif principal de ce travail est divisé en trois objectifs, qui sont mis en œuvre étape par étape.

Le premier objectif est de concevoir un réseau neuronal convolutif (CNN) hautes performances, spécifique au patient, avec une taille de modèle minuscule pour une prédiction efficace des crises, qui aide à soulager l'anxiété du patient (faible taux de fausses prédictions (FPR)) et permet des mesures préventives ou des interventions médicamenteuses (haute sensibilité). Pour atteindre le premier objectif, inspiré de CNN unidimensionnel (1D CNN) et VGGNet, un neurone convolutionnel empilé unidimensionnel réseau (1DSCNN) est proposé pour prédire les crises d'épilepsie. Ce modèle est conçu avec une petite taille très compétitive adaptée aux dispositifs biomédicaux portables. Par rapport aux méthodes de pointe, le 1DSCNN proposé obtient les meilleures performances avec la taille de modèle la plus petite sur l'ensemble de données de l'American Epilepsy Society Seizure Prediction Challenge (AES). De plus, un schéma de quantification préliminaire est également appliqué pour évaluer l'impact de différentes largeurs de bits sur les performances du modèle, facilitant ainsi son déploiement dans les dispositifs biomédicaux portables.

Le deuxième objectif est d'explorer plus avant un d'apprentissage profond (DL) économe en énergie et spécifique au patient pour prédire avec précision les crises d'épilepsie, en abordant le compromis entre performances et consommation d'énergie dans les dispositifs médicaux portables aux ressources limitées. Pour atteindre le deuxième objectif, premièrement, notre 1DSCNN proposé dans le premier objectif est ensuite évalué pour démontrer sa généralisation sur les ensembles de données AES et EEG du cuir chevelu du Boston Children's Hospital-MIT (CHB-MIT), qui dépasse l'état de l'art. -méthodes artistiques en termes de sensibilité, FPR, aire sous la courbe des caractéristiques de fonctionnement du récepteur (ROC) (AUC), taille du modèle et consommation d'énergie. Ensuite, un schéma de quantification à précision fixe et convivial a été mis en œuvre sur les deux ensembles de données, ce qui permet d'obtenir une excellente efficacité énergétique avec des pertes de performances minimales. Enfin, une méthode de recherche à précision mixte, presque optimale et de faible complexité, est également suggérée, ce qui pourrait potentiellement améliorer les performances du modèle pour les sujets difficiles à prédire souffrant de crises d'épilepsie.

Le troisième objectif est de développer un CNN performant, interprétable et peu gourmand en énergie pour prévoir efficacement les crises d'épilepsie, qui se veut transparent et compréhensible pour les professionnels de santé. Pour atteindre le troisième objectif, un mécanisme d'attention spectral-spatial empilé 1D CNN ($S^3$1DCNN) axé sur l'attention est proposé pour prévoir les crises, mettant en vedette sa capacité interprétable à analyser les enregistrements spatio-temporels non stationnaires EEG pour localiser avec précision les régions d'épilepsie. début. De plus, le $S^3$1DCNN proposé surpasse les méthodes de pointe concernant AUC, la taille du modèle et la consommation d'énergie sur l'ensemble de données AES, ce qui montre son excellent potentiel pour les dispositifs portables biomédicaux de faible puissance.

En résumé, cette thèse explore et met en œuvre des approches pratiques pour les dispositifs médicaux portables et implantables à ressources limitées afin de fournir aux patients épileptiques une prédiction fiable des crises. Cela améliore leur capacité à gérer et à anticiper efficacement de telles crises.

ABSTRACT

Seizure prediction is vital for people with drug-resistant epilepsy, who may not respond well to antiepileptic drugs (AEDs). For these patients, it significantly improves safety by allowing individuals to take precautionary measures or alternative interventions against potential injuries associated with unexpected seizures. In addition, seizure prediction helps minimize the reliance on medications that can have severe side effects. Many AEDs come with various side effects, which can affect patient quality of life. By enabling better timing and reducing medication dosage through precise prediction, patients can experience fewer side effects while maintaining effective seizure control. Thus, wearable or implantable devices are critical in epilepsy management because they provide continuous and real-time electroencephalogram (EEG) signal monitoring. This capability allows for detecting precursory seizure patterns, enabling early warnings and interventions that can improve patient autonomy and quality of life.

This thesis focuses on a robust epileptic seizure prediction engine tailored for wearable medical devices to explore practical approaches to help epileptic patients obtain seizure prediction information in time. This work's core aim is to develop high-performance, energy-efficient, and interpretable methods for seizure prediction in resource-limited wearable devices that can potentially be suitable for daily use. The main challenge of this work lies in the trade-off between performance and energy consumption, highlighting an ongoing dilemma in resource-limited wearable medical devices. Furthermore, the core aim of this work is divided into three objectives, which are implemented step by step.

The first objective is to design a high-performance, patient-specific convolutional neural network (CNN) with a tiny model size for effective seizure prediction, which helps alleviate patient anxiety (low false prediction rate (FPR)) and enable preventive measures or medication interventions (high sensitivity). To achieve the first objective, inspired by one-dimensional CNN (1D CNN) and VGGNet, a one-dimensional stacked convolutional neural network (1DSCNN) is proposed to predict epilepsy seizures. This model is designed with a very competitive small size suitable for wearable biomedical devices. Compared to state-of-the-art methods, the proposed 1DSCNN achieves the best performance with the smallest model size on the American Epilepsy Society Seizure Prediction Challenge (AES) dataset. In addition, a preliminary quantization scheme is also applied to evaluate the impact of various bit widths on model performance, facilitating its deployment in wearable biomedical devices.

The second objective is to explore an energy-efficient, patient-specific deep learning (DL) algorithm for effectively predicting epileptic seizures, addressing the trade-off between performance and energy consumption in resource-limited wearable medical devices. To achieve the second objective, firstly, our proposed 1DSCNN in the first objective is further evaluated to demonstrate its generalization on the AES and Boston Children's Hospital-MIT scalp EEG (CHB-MIT) datasets, which surpasses state-of-the-art methods in terms of sensitivity, FPR, area under the receiver operating characteristic (ROC) curve (AUC), model size and energy consumption. Then, a hardware-friendly, fixed-precision quantization scheme has been implemented on the two datasets, resulting in excellent energy efficiency with minimal performance losses. Finally, a near-optimal, low-complexity mixed-precision search method is also suggested, which has the potential to improve model performance for hard-to-predict subjects with epileptic seizures.

The third objective is to develop a high-performance, interpretable CNN with minor energy consumption for accurately forecasting epileptic seizures, which is intended to be transparent and understandable to healthcare professionals. To achieve the third objective, an attention-driven stacked spectral-spatial attention 1D CNN ($S^3$1DCNN) is proposed to forecast seizures, featuring its interpretable ability to analyze spatio-temporal non-stationary EEG recordings for precisely localizing regions of epilepsy onset. In addition, the proposed $S^3$1DCNN outperforms state-of-the-art methods regarding AUC, model size, and energy consumption on the AES dataset, which shows its excellent potential for low-power biomedical wearable devices.

To summarize, this thesis explores and implements practical approaches for resource-limited wearable and implantable medical devices to provide epilepsy patients with reliable seizure prediction, enhancing their ability to manage and anticipate seizures effectively.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ILAE	International League Against Epilepsy
INN	International Nonproprietary Names
EEG	Electroencephalogram
CT	Computerized Tomography
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
sEEG	Scalp EEG
iEEG	Intracranial EEG
AEDs	Antiepileptic Drugs
SPH	Seizure Prediction Horizon
DL	Deep Learning
ML	Machine Learning
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
AES	American Epilepsy Society Seizure Prediction Challenge
CHB-MIT	Boston Children's Hospital-MIT Scalp EEG
STFT	Short-Time Fourier Transform
DWT	Discrete Wavelet Transform
SMOTE	Synthetic Minority Over-Sampling Technique
NMF	Nonnegative Matrix Factorization
CWT	Continuous Wavelet Transform
GAN	Generative Adversarial Network
PE	Permutation Entropy
MLPs	Multi-Layer Perceptrons
DCNN	Deep Convolutional Neural Network
Bi-LSTM	Bidirectional Long Short-Term Memory Network
DCAE	Deep Convolutional Autoencoder
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory Network
FIR	Finite Impulse Response
NAS	Neural Architecture Search
SVM	Support Vector Machines
DTs	Decision Trees

kNN	k-Nearest Neighbors
PCA	Principal Component Analysis
RBF	Radial Basis Function
FPR	False Prediction Rate
CSP	Common Spatial Pattern
GA	Genetic Algorithm
SOM	Self-Organizing Maps
SDCN	Semi-Dilated Convolutional Neural Networks
LDA	Linear Discriminant Analysis
1DSCNN	One-Dimensional Stacked Convolutional Neural Network
2D CNN	Two-Dimensional CNN
1D CNN	One-Dimensional CNN
ReLU	Rectified Linear Unit
Tanh	Hyperbolic Tangent
STE	Straight-Through Estimators
FP	Full Precision
Mv-CGRN	Multi-View Convolutional Gated Recurrent Network
BNLSTM	Batch Normalization Long Short-Term Memory Networks
CASA	Channel and Spatial Attention
MAC	Multiply-Accumulate
SRAM	Static Random-Access Memory
3D CNN	Three-Dimensional CNN
BN	Batch Normalization
$S^3$1DCNN	Stacked Spectral-Spatial Attention 1D CNN
Grad-CAM	Gradient-Weighted Class Activation Mapping
Adam	Adaptive Moment Estimation
RMSprop	Root Mean Square Propagation
QAT	Quantization-Aware Training
DNN	Deep Neural Network

CHAPTER 1 INTRODUCTION

1.1 Epilepsy and Epileptic Seizures

1.1.1 Epilepsy Overview

With an incidence of fifty new diagnoses every 100,000 people annually and an estimated occurrence of 0.5% in the population, epilepsy is one of the most prevalent neurological disorders marked by a persistent predisposition to generate epileptic seizures and by the neurobiological, cognitive, psychological, and social consequences of this condition [5]. Affecting more than 65 million people worldwide, epilepsy is characterized by the recurrent occurrence of seizures in the brain cortex [6]. Seizure manifestations are brief episodes of involuntary movement that can involve a part of the body or the entire body and are sometimes accompanied by loss of consciousness [7], significantly affecting the quality of life of patients and their social participation.

According to the definition provided by International League Against Epilepsy (ILAE), epilepsy is diagnosed under any of the following conditions: 1) At least two unprovoked or reflex seizures occurring more than 24 hours apart; 2) A single unprovoked or reflex seizure with a subsequent probability of further seizures exceeding 60% over the following 10 years; 3) Identification of an epileptic syndrome. These epileptic seizures manifest in various forms and occur at varying frequencies. The causes of epilepsy are multifaceted, including genetic disorders, brain structure abnormalities, immune system abnormalities, metabolism changes, infections, brain trauma, brain tumors, and stroke, highlighting the complexity of this condition [8]. Despite the diversity of causes, the commonality shared across many epilepsy cases is the profound unpredictability of seizure events. This unpredictability poses significant safety risks and psychosocial burdens on people living with epilepsy [9], as unexpected seizures can lead to short instances of uncontrollable movement or even loss of consciousness, while fear of potential seizures can limit employment opportunities, social interactions, and personal independence [10].

1.1.2 Epileptic Seizures

Seizures occur due to abnormal electrical activity in the brain, altering bodily functions. They manifest in various forms, potentially causing convulsions, muscle spasms, transient or sustained loss of consciousness, unusual sensations and emotions, and atypical behaviors. Typically, epileptic seizures are categorized into two broad types according to their various points of origin in the brain, as shown in Figure 1.1.

Focal Seizures

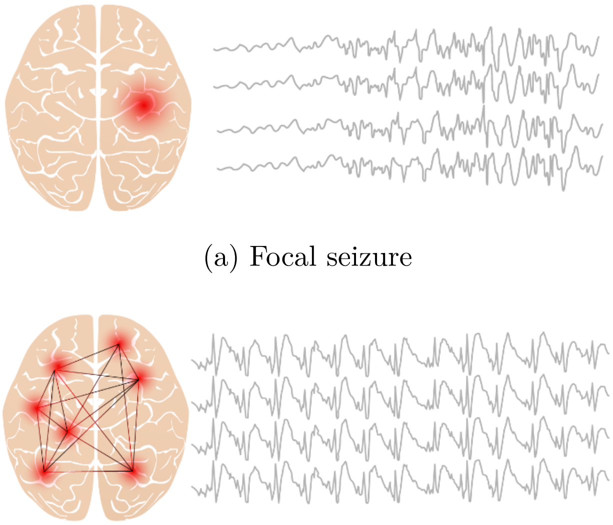
Focal epileptic seizures, which begin as single-vision, multifocal, or encompass a cerebral hemisphere, originate in specific areas of the brain. The types of focal seizures are: 1) Motor focal seizures; 2) Sensory focal seizures; 3) Autonomic focal seizures; 4) Psychological focal seizures. These seizures affect approximately 60% of people diagnosed with epilepsy. The symptoms experienced by people with focal epileptic seizures vary widely, depending on both the origin and the extent of spread within the brain. These symptoms can range from simple sensations such as finger tingling to complex perceptual disturbances, including hallucinations and sensory distortions (olfactory, visual, and auditory). Focal seizures may preserve or alter the state of consciousness because these seizures have the potential to evolve into generalized seizures.

Generalized Seizures

Generalized epileptic seizures initiate simultaneously on both sides of the brain. The types of generalized seizures are: 1) Absence (petit-mal); 2) Atonic; 3) Tonic; 4) Clonic; 5) Myoclonic; 6) Tonic-clonic (grand-mal). Manifestations of these seizures include bilateral movements such as brief jerks, muscle stiffness, or convulsions, as well as episodes of unresponsiveness, known as absence seizures. Typically, this type of seizure is associated with transient alterations or a complete loss of consciousness. It is essential to recognize that individuals can experience multiple types of seizures, oscillating between focal seizures and generalized seizures.

1.1.3 Diagnosis of Epilepsy

Diagnostic procedures are performed to determine if a person has epilepsy and, if applicable, the kind of seizures they may have. The healthcare team uses imaging and monitoring examinations such as EEG, computerized tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET) to comprehend the effect of epilepsy on a patient. Neuroimaging techniques such as CT, MRI, and PET are employed when the origin of epileptic seizures is unknown, which help in identifying structural and functional abnormalities in the brain that may be underlying the epilepsy. However, EEG stands out as a more practical choice over neuroimaging techniques due to its adaptability and efficiency in continuous monitoring, especially for wearable and implantable devices [11]. Furthermore,



(a) Focal seizure

(b) Generalized seizure

Figure 1.1 Types of epileptic seizures

the high temporal resolution of EEG makes it superior for detecting the precise onset and duration of seizures, which is crucial for the effective diagnosis and treatment of epilepsy [6]. Here are two main types of EEG: scalp EEG (sEEG) and intracranial EEG (iEEG).

sEEG

sEEG is a non-invasive diagnostic technique that involves the placement of electrodes on the scalp to detect and record the brain's electrical activity. This technique is fundamental in studying and diagnosing neurological conditions, especially epilepsy [12]. sEEG is instrumental in identifying abnormal electrical activity that can indicate epileptic seizures, brain dysfunctions, and other neurological disorders. The international 10-20 system, as shown in Figure 1.2, uses multiple electrodes placed according to standardized locations on the scalp to provide comprehensive coverage of the cerebral cortex, thus maximizing the diagnostic yield. The captured electrical signals are amplified and displayed, allowing clinicians to analyze brain waves' frequency, amplitude, and spatial distribution [13]. This analysis helps diagnose conditions, guide treatment decisions, and assess brain function in real time. The characteristics of sEEG, such as its accessibility, non-invasive nature, and ability to effectively monitor in real-time, make it an essential tool in clinical neurology.

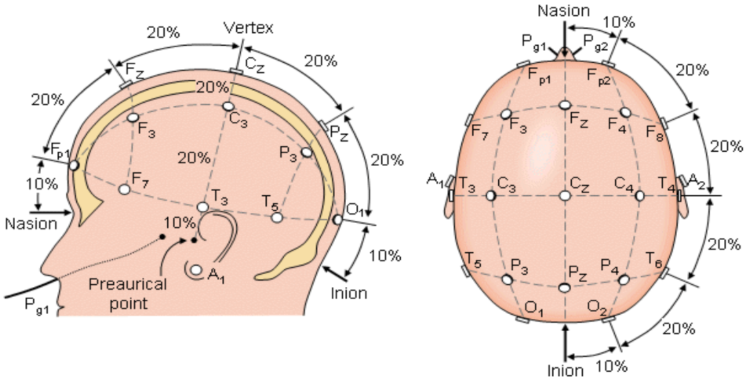


Figure 1.2 The international 10-20 system of sEEG electrode placement [1]

iEEG

iEEG is an advanced neurophysiological monitoring technique used predominantly in the pre-surgical evaluation of epilepsy patients who do not respond to medical therapy. This invasive method involves placing electrodes directly onto the brain's surface or within the brain tissue, as demonstrated in Figure 1.3, to record electrical activity. iEEG is crucial for localizing seizure foci with high spatial resolution, particularly in complex cases where sEEG is insufficient [14]. By providing a detailed map of brain electrical activity, iEEG helps to precisely identify the areas responsible for seizure generation, which is essential for planning curative surgical interventions [15]. Unlike sEEG, iEEG can bypass the skull's and scalp's blurring effects, offering direct recordings of neuronal discharges and, thus, more accurate and detailed information. In addition, during iEEG examinations, the placement of electrodes is tailored to patient-specific needs. This method is invaluable for determining brain functional areas, such as those involved in language and motor functions, thereby minimizing the risks of neurological deficits post-surgery.

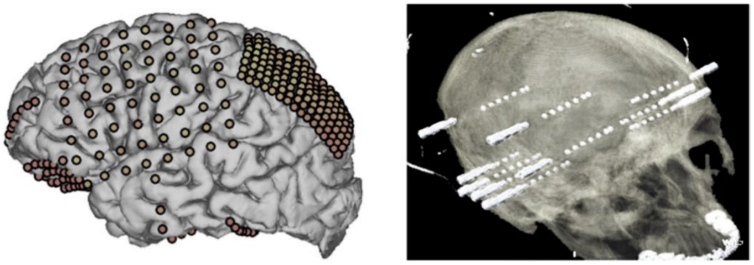


Figure 1.3 Two methods of iEEG: subdural electrodes (left) and depth electrodes (right) [2]

1.1.4 Medical Therapy

Drug Resistance

Seizure control with medications is shown in Figure 1.4. Despite advances in medical science, 32% of epilepsy patients do not achieve seizure control with current antiepileptic drugs (AEDs). Approximately 50% of patients become seizure-free after the first trial of AEDs. An additional 15% find success with the second drug trial, while only 4% achieve seizure control with the third medication. However, for the remaining patients, epilepsy continues to be uncontrolled. The limitations of current treatments underscore the urgent need for innovative approaches in epilepsy management, particularly in the area of seizure prediction and proactive therapeutic strategies.

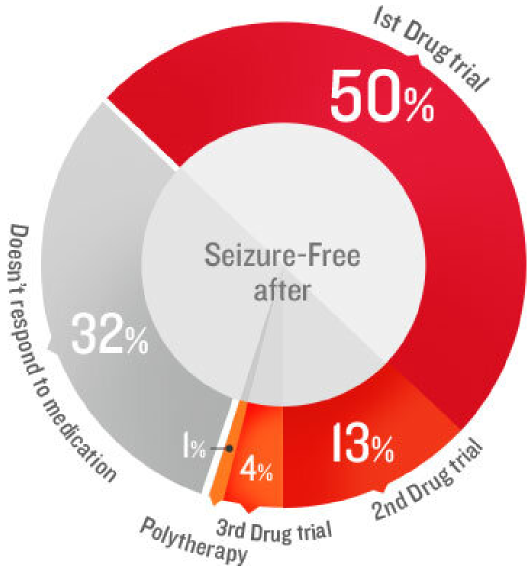


Figure 1.4 Seizure control with medications [3]

Side Effects

Many AEDs are metabolized hepatically, which accounts for numerous potential drug interactions. The risk of these interactions is compounded by the fact that certain AEDs can alter hepatic metabolism. This alteration occurs through enzyme induction, as seen with drugs such as phenobarbital, carbamazepine, phenytoin, and oxcarbazepine, or through enzyme inhibition, as with valproic acid, felbamate, and topiramate. The numerous drug interactions arising from the hepatic metabolism of AEDs contribute to patient treatment response variability. There is also inter-individual variability, as demonstrated by the resistance to treatment observed in 32% of patients. Also, all AEDs have cerebral adverse effects, which are associated with their psychotropic characteristics. AEDs are associated with numerous and frequent side effects, as detailed in Table 1.1. These side effects underscore the complexity of managing treatment regimens in epilepsy care.

Table 1.1 Side effects of different AEDs

INN Drug	**Idiosyncratic Acute Adverse Reactions**	**Acute Dose-dependent Adverse Reactions**	**Chronic Adverse Reactions**
Phenobarbital	Rash, Stevens Johnson, liver damage, agranulocytosis, thrombocytopenia, induced lupus, rheumatism	Drowsiness	Behavioural disorders, cognitive impairment
Carbamazepine	Leukopenia/agranulocytosis, Stevens Johnson, morbilliform rash, cytolytic liver involvement, aplastic anemia	Drowsiness, dizziness, ataxia, nausea, vomiting, blurred vision, diplopia, cardiac conduction disorders	Hyponatremia
Phenytoin	Blood dyscrasia, induced lupus, rash, Stevens Johnson, hepatotoxicity	Ataxia, nystagmus, nausea, vomiting	Acne, hirsutism,thick visage, gingival enlargement, pressure, confusion, nemia, megaloblastic
Oxcarbazepine	Rash	Sensations, dizziness, diplopia, dyskinesia, ataxia, headache	Hyponatremia
Valproic acid	Acute pancreatitis, cytolytic hepatitis, thrombocytopenia, encephalopathy, leukopenia	Dyspepsia, nausea, vomiting	Tremor, weight gain, alopecia, peripheral edema
Felbamate	Aplastic anemia, cytolytic hepatitis	Alterations in cognitive functions, dizziness	Impairments in cognitive functions, anorecco effect with weight loss
Topiramate	Secondary angle-closure glaucoma	Drowsiness, dizziness, ataxia, anemia, leukopenia	Alopecia, kidney stones, psychomotor slowdown, cognitive impairment

international nonproprietary names (INN).

1.2 Problem Formulation

Epileptic seizures pose important challenges in healthcare, and the ability to predict them could significantly improve patients' quality of life when they cannot be prevented effectively by medication alone. This thesis focuses on enhancing the performance of classification methods and algorithms that can be used to predict preictal and interictal states. It focuses on improving their energy efficiency for platforms that could perform real-time inference on wearable devices. Although the ultimate goal of this research is to enable implementation on wearable platforms, this work focuses on algorithms without delving into any particular embedded platform. The energy efficiency of these algorithms is examined as a step toward developing solutions that perform well on resource-limited platforms. A key contribution of this thesis lies in its focus on models with a small memory footprint, which is essential for successful implementation on low-power, embedded systems. This thesis emphasizes the foundational principles of energy consumption in low-power wearable systems, particularly for application-specific accelerators such as deep neural network (DNN) accelerators, where energy usage is dominated by data movement and computation during inference.

1.2.1 Seizure Prediction

Epileptic seizure prediction holds utmost significance, particularly from the perspectives of drug resistance and drug-induced side effects. On the one hand, approximately one-third of epilepsy patients exhibit refractory responses to AEDs, which highlights the need for alternative intervention strategies, such as the prediction of seizures that could proactively warn patients about upcoming seizures. On the other hand, the side effects associated with AEDs can be debilitating, ranging from mild dizziness to severe depression and cognitive impairment. Therefore, seizure prediction could play a crucial role by optimizing the timing and dosage of the medication administered, improving therapeutic outcomes while reducing adverse effects.

Seizure prediction is commonly viewed as a binary classification task between preictal and non-preictal states. The preictal state refers to the period before a seizure onset. In contrast, the non-preictal state consists of three distinct phases: interictal (seizure-free intervals), ictal (seizure periods), and postictal (post-seizure periods). Seizure prediction horizon (SPH) offers a window for implementing medical treatments or preventive strategies during the transition from preictal to ictal states. An illustration of these states is presented in Figure 1.5. Predictions made during the preictal period are considered true, while those made during non-preictal periods are deemed false. Epileptologists can typically identify the ictal and postictal phases by analysis of EEG recordings and patient video monitoring during hospitalization [16].

However, the main challenge in seizure prediction lies in the accurate classification of preictal versus interictal states, one of the central objectives of this study. Current technology struggles with high rates of false positives and false negatives, which can significantly disrupt the daily lives of patients and the confidence in the technology [17]. The complexity of this task comes from the subtle variations in EEG that distinguish these states, requiring advanced analytical methods and robust predictive models, such as promising deep learning (DL) techniques, to improve reliability and efficacy.

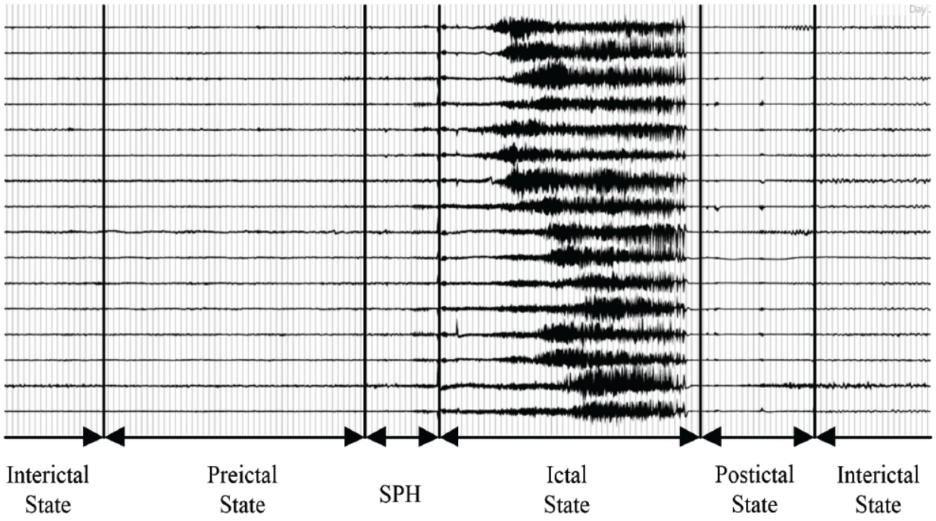


Figure 1.5 Epileptic brain states: interictal, preictal, ictal, SPH, and postictal

1.2.2 Energy Efficiency

The advent of wearable and implantable devices for epileptic seizure prediction marks a significant advancement in treating epilepsy [18]. However, the trade-off between performance and energy consumption is evident when contrasting traditional machine learning (ML) algorithms with DL counterparts [19], as depicted in Figure 1.6. Traditional ML-based classifiers offer the advantage of low energy consumption, typically requiring less than 100 μJ per inference, paired with a model size of less than 1 kB, thus improving the energy efficiency of wearable and implantable devices. However, this comes at the expense of AUC, often falling below 0.95. On the other hand, DL-based classifiers, while surpassing AUC of 0.95, indicate a considerable increase in energy demand, exceeding 10 mJ per inference, and require a larger model size of over 50 kB. Such requirements significantly burden finite energy resources, particularly for implants where battery life is a critical constraint. Thus, this pivotal constraint highlights the need for algorithms that balance high AUC and low energy consumption for DL-based classification. Based on these considerations, this thesis aims to propose methods that allow accurate and timely prediction of imminent epileptic seizures with low complexity models that can be executed in real-time on hardware platforms consuming low energy.

1.3 Claimed Contributions

The main goal of this thesis is to develop a robust prediction engine for epileptic seizures based on the DL algorithm. To achieve this, we divide the main goal into three secondary objectives, focusing on the previously mentioned aims through in-depth research.

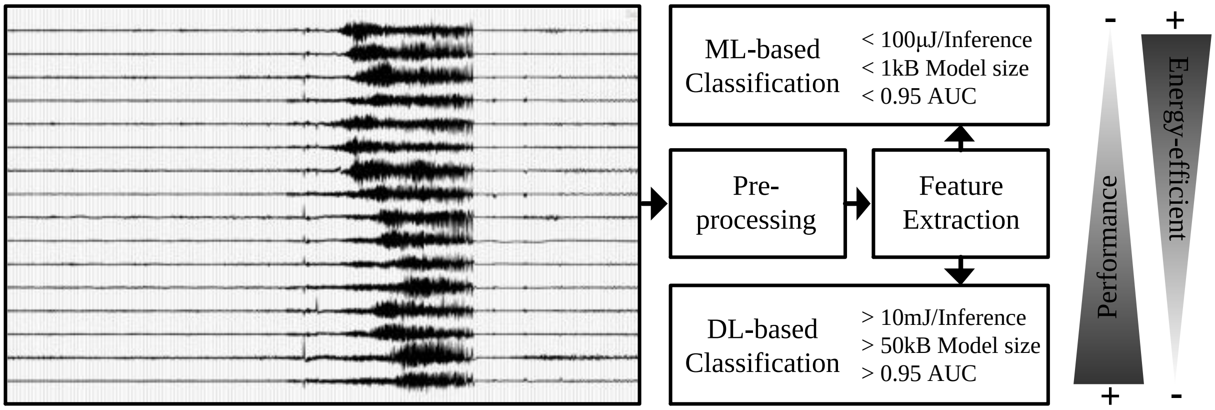


Figure 1.6 The trade-off of epileptic seizure prediction between ML-based classification and DL-based classification

Objective 1: To design a high-performance, patient-specific DL algorithm based on a tiny neural network model for accurate epileptic seizure prediction, which helps reduce patient anxiety (low false prediction rate (FPR)) and facilitates precautionary actions or medication interventions (high sensitivity). In pursuit of the first objective, a one-dimensional stacked convolutional neural network (1DSCNN) is proposed to predict epileptic seizures. This model features a highly efficient and small memory footprint, which makes it well-suited for embedding into wearable biomedical devices. The proposed 1DSCNN outperforms leading methods while maintaining the smallest model size on long-term continuous iEEG recordings. Furthermore, a preliminary quantization strategy is implemented to assess how different bit widths affect model performance, thus supporting its application within wearable biomedical devices. This objective was met and led to the publication of Article 1 provided in Chapter 4.

- Article 1: **Y. Zhang**, Y. Savaria, S. Zhao, G. Mordido, M. Sawan, and F. Leduc-Primeau, "Tiny cnn for seizure prediction in wearable biomedical devices," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 1306-1309, doi: 10.1109/EMBC48229.2022.9872006. (Oral presentation)

Objective 2: To explore an energy-efficient, patient-specific DL algorithm for effectively predicting epileptic seizures, which can balance performance and energy consumption in wearable medical devices with limited resources. In pursuit of the second objective, initially, the proposed 1DSCNN implemented while tackling the first objective is further assessed to show its ability to generalize on long-term continuous iEEG and sEEG recordings, which outperforms state-of-the-art methods regarding sensitivity, FPR, AUC, model size, and energy consumption. Subsequently, a hardware-efficient, fixed-precision quantization approach was applied and validated on the abovementioned recordings, achieving remarkable energy efficiency with negligible performance degradation. Finally, a near-optimal, low-complexity mixed-precision search strategy is also proposed, which could enhance model performance for patients with epileptic seizures that are difficult to anticipate. This objective was met, and the research results are reported as Article 2, in Chapter 5.

- Article 2: **Y. Zhang**, Y. Savaria, M. Sawan, and F. Leduc-Primeau, “Tiny Neural Network for Epileptic Seizure Forecasting in Wearable Devices,” *IEEE Transactions on Biomedical Engineering*, submitted, 2024/03/20.

Objective 3: To devise a high-performance, interpretable DL algorithm with acceptably low energy consumption that can accurately forecast epileptic seizures, aiming to be transparent and understandable to healthcare professionals. In pursuit of the third objective, an attention-based stacked spectral-spatial attention one-dimensional convolutional neural network (CNN) (1D CNN) ($S^3$1DCNN) is designed to predict seizures, highlighting its capacity to interpret spatiotemporal non-stationary iEEG recordings for accurately identifying epilepsy onset regions. Furthermore, the proposed $S^3$1DCNN surpasses state-of-the-art approaches regarding AUC, model size, and energy consumption, demonstrating its remarkable suitability for low-energy biomedical wearable devices. The objective was met, and the research results are reported in Article 3, as detailed in Chapter 6.

- Article 3: **Y. Zhang**, Y. Savaria, M. Sawan, and F. Leduc-Primeau, “$S^3$1DCNN: A compact stacked spectral-spatial attention 1DCNN for seizure prediction with wearables,” in *2024 22st IEEE Interregional NEWCAS Conference (NEWCAS)*. IEEE, 2024, pp. 278-282, doi: 10.1109/NewCAS58973.2024.10666297. (Oral presentation)

1.4 Thesis Outline

This chapter reviews epilepsy, epileptic seizures, EEG types, and problems with related existing medical therapy. It illustrates the research aims and highlights the key contributions of this thesis.

The subsequent chapters are organized as follows. Chapter 2 provides a detailed literature review on epileptic seizure prediction. Chapter 3 provides fundamental theory and methodology supporting the subsequent chapters. Chapter 4, Chapter 5 and Chapter 6 include published and submitted papers that aim to develop a robust prediction engine of epileptic seizures based on DL algorithm. Chapter 7 expands on the findings of Chapter 4 and provides further analysis of the seizure prediction method discussed in Chapter 6. Chapter 8 offers an overall thesis discussion, considering the methods, contributions, results, and significance illustrated in the previous chapters. This discussion also proposes a comparison of the results obtained while tackling the three main objectives of the Thesis. Chapter 9 outlines the main findings of this work, then engages in a general discussion of limitations and proposes future research.

CHAPTER 2 LITERATURE REVIEW

2.1 History of Seizure Prediction

Epilepsy ranks as the second most common neurological disorder, surpassed only by stroke. Consequently, the quest for effective treatments has gained significant attention from medical researchers and technology experts. Although AEDs and surgical interventions exist, they fail to offer a cure for all patients and often induce unexpected side effects. Seizure prediction emerges as a promising direction for advancing epilepsy treatment [20]. It is essential to review the development of seizure prediction research to understand the practical problems in this domain better. Figure 2.1 provides a historical overview of developments in seizure prediction.

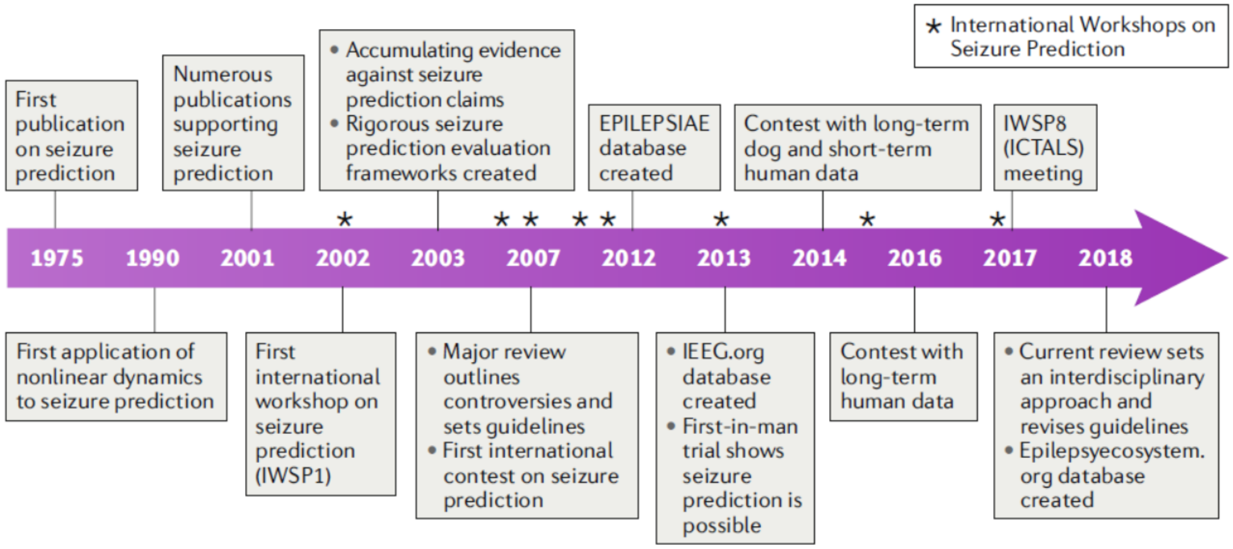

★ International Workshops on Seizure Prediction
First publication on seizure prediction
Numerous publications supporting seizure prediction
• Accumulating evidence against seizure prediction claims
• Rigorous seizure prediction evaluation frameworks created
EPILEPSIAE database created
Contest with long-term dog and short-term human data
IWSP8 (ICTALS) meeting
1975
1990
2001
2002
2003
2007
2012
2013
2014
2016
2017
2018
First application of nonlinear dynamics to seizure prediction
First international workshop on seizure prediction (IWSP1)
• Major review outlines controversies and sets guidelines
• First international contest on seizure prediction
• IEEG.org database created
• First-in-man trial shows seizure prediction is possible
Contest with long-term human data
• Current review sets an interdisciplinary approach and revises guidelines
• Epilepsyecosystem.org database created

Figure 2.1 The evolution of seizure prediction [4]

The establishment of the seizure prediction domain dates back to the 1970s and after those early investigations on the predictability of seizures. Multiple groups attempted to employ linear methods to identify seizure precursors in sEEG data from absence seizures [21]. Then, in the 1990s, the advent of non-linear systems was successfully applied to seizure prediction for better matching the non-linear characteristics of EEG data [22]. Martinerie et al. [23] transformed iEEG recordings into trajectories within a phase space, utilizing correction density to indicate a reduction in spatio-temporal complexity during the preictal phase. This period also saw the categorization of EEG patterns into preictal, ictal, and interictal phases, with the preictal phase being used to predict seizures. As the millennium began, a significant amount of research accumulated in this area, prompting scholars from diverse fields to pay more attention to predicting seizures. In 2002, multiple epilepsy centers organized the first international workshop on seizure prediction, which stimulated further research in this field [24]. Then in 2003, Drogenlen et al. used Kolmogorov entropy to predict epileptic seizures between 2 and 40 minutes before seizure onset [25]. Meanwhile, Mormann et al. identified a reduction in phase synchronization across EEG channels before a seizure [26]. Although various approaches have been developed to address seizure prediction problems, the variability in databases used by different research groups and variations in prediction timing for seizure onset and the duration of the presumed preictal phase complicates these efforts. To this end, the first seizure prediction competition was held in 2007 and was associated with the third international workshop on seizure prediction.

In 2009, a second competition followed as part of the fourth international workshop on seizure prediction. These competitions provided participants with continuous iEEG recordings from three patients, but the algorithms failed to meet the performance expectations. In 2012 and 2013, the EPILEPSIAE [27] and IEEG.org[1] databases were established, respectively. Meanwhile, it should be noted that in 2013, the first-in-man trial demonstrated that seizure prediction was indeed possible [17]. In 2014, the American Epilepsy Society Seizure Prediction Challenge[2] involved the analysis of short-term human iEEG and long-term canine iEEG with epilepsy. Similarly, the 2016 Melbourne University AES/MathWorks/NIH Seizure Prediction contest[3] used long-term iEEG recordings containing 1139 seizure clips. Despite these efforts, the optimal features and algorithms for seizure prediction remain unclear. The complexity of submitted algorithms has made determining the most effective approaches challenging. Thus, in 2018, Epilepsyecosystem.org[4] was created to examine these intricate algorithms and determine effective solutions.

2.2 Public Databases

2.2.1 Bonn Dataset

The Bonn dataset[5] is commonly used in epilepsy research, particularly in evaluating seizure detection or prediction algorithms. The data set is sampled at a frequency of 173.61 Hz, comprising 500 frames, each lasting 23.6 seconds [28]. It contains EEG recordings divided

[1] https://www.ieeg.org/

[2] https://www.kaggle.com/competitions/seizure-prediction/

[3] https://www.kaggle.com/c/melbourne-university-seizure-prediction

[4] https://www.epilepsyecosystem.org/

[5] https://www.ukbonn.de/epileptologie/arbeitsgruppen/ag-lehnertz-neurophysik/downloads/

into five subsets (A, B, C, D, E), where subsets A and B are from sEEG signals of healthy individuals. In contrast, subsets C, D, and E contain iEEG signals from epilepsy patients in interictal intervals (C and D) and ictal intervals (E).

2.2.2 Freiburg Hospital Dataset

The Freiburg Hospital dataset[6] is built by performing invasive presurgical monitoring in the center of the Freiburg University Hospital. It contains iEEG recordings from 21 patients with intractable focal epilepsy. These iEEG recordings were acquired using 6 electrodes (3 focal, 3 extra focal), sampled at 256Hz. The dataset includes 24 hours of interictal recordings and at least 50 minutes of preictal recordings for each patient. Overall, the dataset contains 21 patients with 88 seizures and 582 hours of iEEG data.

2.2.3 CHB-MIT Dataset

The Boston Children's Hospital-MIT scalp EEG (CHB-MIT) dataset[7] is a widely used dataset, featuring sEEG recordings from 23 pediatric patients (5 males ages 3–22, 17 females ages 1.5–19, and one unknown) grouped into 24 cases with medically intractable focal epilepsy [29]. These sEEG recordings are collected by the international 10-20 system. The dataset contains 969 hours of non-invasive recordings [30], sampled at 256Hz, and 198 seizures.

2.2.4 AES Dataset

The American Epilepsy Society Seizure Prediction Challenge (AES) dataset[8] consists of iEEG recordings from five dogs and two human patients with 48 seizures and a total recording of 627 hours [31]. These iEEG recordings are collected using an ambulatory monitoring system. Four dogs in this dataset are monitored using 16 subdural electrodes each, while the remaining dog is monitored with 15 electrodes, with a sampling rate of 400 Hz. Recordings for one patient are required using 15 subdural electrodes, whereas another patient is recorded with 24 electrodes, sampled at 5000 Hz. These recordings cover periods ranging from several months to a full year.

[6] https://www.epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database/

[7] https://physionet.org/content/chbmit/1.0.0/

[8] https://www.kaggle.com/c/seizure-prediction/

2.2.5 Melbourne University Dataset

The Melbourne University dataset[9] can be accessed through the Melbourne University AES/MathWorks/NIH Seizure Prediction contest. It includes iEEG recordings from three patients with 211 seizures and total recordings of 442 days [32], captured using an ambulatory monitoring system. The monitoring system employs 16-channel electrodes placed on the surface of the cerebral cortex, sampled at 400 Hz.

2.2.6 Databases Comparison

The comparison of different public datasases is detailed in Table 2.1. The database recordings have progressively increased over time, enabling a more detailed examination of complex algorithms to determine effective solutions.

Table 2.1 Comparison of recent public databases

Database	**Published Year**	**Recording Type**	**No. of Subjects**	**No. of Channels**	**No. of Seizures**	**Sampling Rate (Hz)**	**Total Hours**
Bonn	2001	sEEG/iEEG	5	1	-	173.61	3.28
Freiburg	2003	iEEG	21	128	88	256	582
CHB-MIT	2010	sEEG	24	23, 24, 26	198	256	969
AES	2014	iEEG	7	16, 15, 24	48	400, 5000	627
Melbourne	2016	iEEG	3	16	211	400	10608

2.3 Feature Extraction

Feature extraction typically involves the transformation of raw time-series data into a format more easily analyzed by non-end-to-end algorithms. This process includes preprocessing methods such as time-frequency analysis, which breaks down the EEG signals into time and frequency components, providing a comprehensive view of the signal characteristics essential to differentiate between preictal and interictal states. In the meantime, as promising end-to-end DL algorithms arise, direct and automatic feature extraction has gained attention for analyzing non-stationary EEG signals. This approach involves feeding raw EEG signals

[9] https://www.kaggle.com/c/seizure-prediction/

directly into the end-to-end network for analysis. It has shown effectiveness in handling the complexities of EEG data and potentially eliminates some traditional preprocessing steps.

2.3.1 Non-End-to-End

The comparison of feature extraction methods by non-end-to-end algorithms is detailed in Table 2.2. Tsiouris et al. [33] filtered EEG signals to remove noise and artifacts and segmented them into 1-minute windows. Then, a variety of features were extracted from EEG signals for seizure prediction, including time-domain features (mean, variance) and frequency-domain features (Fourier and wavelet transforms) extracted from each window. Truong et al. [34] preprocessed EEG signals by removing power line noise and generated more preictal segments by an overlapped sampling technique to balance the imbalanced seizure data. Next, short-time Fourier transform (STFT) was utilized to convert raw EEG signals into a time-frequency matrix, for effective seizure prediction through capturing both temporal and spectral characteristics. Kitano et al. [35] segmented EEG data into non-overlapping 4-second windows. Later, discrete wavelet transform (DWT) was applied to EEG segmentations, followed by counting the number of zero-crossings of detail coefficients at level 1 using the Haar Wavelet function for each segmentation, which focused on capturing transient signal characteristics for differentiating between preictal and interictal states. Stojanovic et al. [36] addressed the class imbalance through synthetic minority over-sampling technique (SMOTE), enhancing the representation of the minority class (preictal state) in the training set. Then nonnegative matrix factorization (NMF) was used to decompose the power spectra of EEG signals into the dominant time and frequency components, which helped in capturing essential information from the EEG signals while eliminating noise and outliers. Hussein et al. [37] empolyed continuous wavelet transform (CWT) to map time-series EEG signals into scalograms, capturing time-frequency information to identify seizure-indicative patterns. Korshunova et al. [38] resampled 10-minute EEG clips to 400 Hz, which were filtered using a band-pass filter between 0.1 and 180 Hz. Each clip was further partitioned into 10 non-overlapping 1-minute frames to prepare the data for feature extraction. Next, the Fourier transform was utilized to acquire a logarithmic transformation of the amplitude spectrum, averaged within specific frequency bands (delta, theta, alpha, beta, low-gamma, high-gamma). Truong et al. [39] used STFT on 28-second windows of EEG signals as a preprocessing step to convert these signals into a time-frequency representation, which facilitates subsequent feature extraction. Then, generative adversarial network (GAN) was employed as an unsupervised technique to extract features, in which the discriminator serves as a feature extractor after training. Yang et al. [40] filtered EEG signals to remove 50 Hz power line interference using a notch filter. Next, permutation entropy (PE) was calculated by a 5-second sliding window without overlap for each EEG channel, which measured the complexity of EEG signal to distinguish between the preictal and interictal states. Usman et al. [41] converted EEG signals from the time domain to the frequency domain using STFT.

Table 2.2 Comparison of feature extraction methods by non-end-to-end algorithms

Method	**Published Year**	**Dataset**	**Feature Extractor**
Hussein et al. [37]	2021	CHB-MIT, AES	CWT
Korshunova et al. [38]	2017	AES	Logarithm of the Fourier amplitude
Kitano et al. [35]	2018	CHB-MIT	DWT, Haar Wavelet function
Usman et al. [41]	2020	CHB-MIT	STFT
Stojanovic et al. [36]	2020	EPILEPSIAE, Melbourne	SMOTE, NMF
Truong et al. [34]	2018	Freiburg, CHB-MIT	STFT
Tsiouris et al. [33]	2018	CHB-MIT	Time-domain (mean, variance), frequency-domain (Fourier, wavelet), cross-correlation, graph features
Truong et al. [39]	2019	CHB-MIT, Freiburg, EPILEPSIAE	STFT, GAN
Yang et al. [40]	2018	Freiburg	PE

EPILEPSIAE: Not available in the public domain, https://epilepsy-database.eu/.

2.3.2 End-to-End

The emergence of promising end-to-end DL algorithms has gained significant attraction with automatic feature extraction for processing raw EEG signals. The comparison of automatic feature extraction methods by end-to-end algorithms is detailed in Table 2.3. Li et al. [42] proposed an end-to-end epilepsy seizure prediction method based on multi-layer perceptrons

(MLPs), which included a denoising layer to remove undesired artifacts from EEG signals, a weighted layer that assigned different weights to each channel based on their significance, and a reduction layer to reduce the input length of EEG signals. Daoud et al. [43] used four distinct DL algorithms to automatically extract features from raw EEG signals. MLPs was directly applied to raw EEG data. Deep convolutional neural network (DCNN) combined with MLPs extracted spatial features from different electrode positions. Bidirectional long short-term memory network (Bi-LSTM) was added to DCNN for capturing both spatial and temporal dynamics, which improved prediction accuracy. Deep convolutional autoencoder (DCAE) combined with Bi-LSTM was utilized to optimize the feature extraction process in a semi-supervised learning setup with transfer learning, aiming at real-time application by reducing computational load. Xu et al. [44] demonstrated a mix of one-dimensional and two-dimensional convolution kernels to process EEG signals and extract features. The one-dimensional kernel was used in the early stages of the network to capture temporal features. In contrast, the two-dimensional kernels in the later stages integrated spatial information across multiple channels for accurate prediction. Wu et al. [45] preprocessed raw EEG signals through finite impulse response (FIR) band-pass filters for specific frequency bands, then utilized an end-to-end long short-term memory network (LSTM) for feature extraction and classification. Zhao et al. [19] employed neural architecture search (NAS) to search energy-efficient CNN for automatic feature extraction, which was tailored for implementation in wearable and implantable devices.

Table 2.3 Comparison of automatic feature extraction methods by end-to-end algorithms

Method	**Published Year**	**Dataset**	**Automatic Feature Extractor**
Li et al. [42]	2023	CHB-MIT, AES	MLPs
Daoud et al. [43]	2019	CHB-MIT	MLPs, DCNN, Bi-LSTM, DCAE
Xu et al. [44]	2020	CHB-MIT, AES	CNN
Wu et al. [45]	2023	CHB-MIT	LSTM
Zhao et al. [19]	2022	CHB-MIT, AES, Melbourne	CNN

2.4 Classification

The classification of seizure prediction is essential for developing effective and personalized epilepsy intervention and management strategies, leveraging both historical data and real-time monitoring to mitigate the effects of seizures on daily living. Classification techniques can be divided into ML-based and DL-based methods. Traditional ML-based classifiers consist of k-nearest neighbors (kNN), decision trees (DTs), and support vector machines (SVM), which have been used to classify EEG data into preictal and interictal states according to extracted features. Recent advances involve DL-based classifiers, including MLPs, CNN, LSTM, DCAE, which are capable of automatically extracting relevant features from raw EEG signals and learning complex temporal and spatial patterns of seizures.

2.4.1 Machine Learning

The comparison of classification methods by machine learning algorithms is as detailed in Table 2.4. Stojanovic et al. [36] introduced NMF to decompose the power spectra of iEEG signals into dominant time and frequency components. To manage the high dimensionality and variability of the signals, linear SVM with L1 regularization was used for classification. The SMOTE addressed the class imbalance inherent in EEG signals. The method was evaluated using the EPILEPSIAE and Melbourne University datasets, achieving high accuracy, sensitivity, and specificity in EPILEPSIAE, with some patients reaching 100% in these metrics. However, the Melbourne University dataset showed lower performance, with an accuracy of around 70% and varying sensitivity and specificity ranges. Despite these variations, the approach provides a computationally efficient and interpretable model suitable for real-time clinical applications. Usman et al. [46] implemented a robust seizure prediction method by preprocessing EEG data through common spatial pattern (CSP) filtering to improve signal-to-noise ratio and employing wavelet transform for denoising. principal component analysis (PCA) was then applied to reduce dimensionality and extract critical features from the EEG signals. Next, these features were classified using SVM to distinguish between preictal and interictal states. Evaluated on the CHB-MIT dataset with 22 subjects involving 84 seizures, the method demonstrated high efficacy with an average sensitivity of 93.1%, successfully identifying the preictal state several minutes before a seizure. Chen et al. [47] developed a seizure prediction method using long-term iEEG data from canines, where features were encoded from 20-second windows into 96-dimensional vectors representing power across different frequency bands. A SVM classifier was trained and then periodically retrained on a highly unbalanced AES dataset to accommodate the non-stationary nature of iEEG signals. Later, a novel post-processing scheme was implemented to integrate predictions from mul-

Table 2.4 Comparison of classification methods by machine learning algorithms

Method	Stojanovic et al. [36]	Usman et al. [46]	Chen et al. [47]	Kitano et al. [35]	Ra et al. [48]	Yang et al. [40]	Qureshi et al. [49]	Costa et al. [50]
Published Year	2020	2019	2021	2018	2021	2018	2023	2024
Used Dataset	EPILEPSIAE, Melbourne University dataset	CHB-MIT dataset	AES dataset	CHB-MIT dataset	CHB-MIT dataset	Freiburg Hospital dataset	CHB-MIT dataset	EPILEPSIAE
Number of Subjects	5 patients, 3 patients	22 patients	4 canines	9 patients	22 patients	19 patients	24 patients	40 patients
Preictal Time	5-minute intervals, 30-second horizon	-	10 to 360 minutes	Up to 30 minutes	-	61.93 minutes	30 minutes	20 to 50 minutes, 5-minute intervals
Feature Extraction Methods	NMF to decompose power spectra	Wavelet transform, CSP	Power in the six Berger frequency bands	DWT, Haar wavelet function	PE	PE using a 5-second sliding window without overlap	DWT	Spectral power, Hjorth parameters, wavelet energy
Feature Selection Methods	-	PCA	Window selection	Wavelet transforms, zero-crossing analysis	PE values based kNN and GA	-	-	Grid search
Classifiers	Linear SVM with L1 regularization	SVM	SVM with a group learning	SOM to identify clusters	SVM	SVM with RBF kernel	SVM with RBF kernel	Logistic regression
Accuracy (%)	97.42, 75.2	-	-	91	74.6	-	94.94	-
Sensitivity (%)	95.2, 69	93.1	84	Up to 98	69.51	94	97.43	-
Specificity (%)	99.4, 78.6	-	-	Up to 88	73.14	-	-	-
FPR	-	-	0.78/day	-	-	0.111/hour	0.138/hour	0.36/hour
Time-in-warning	-	-	0.27	-	-	-	-	-
Seizure Sensitivity	-	-	-	-	-	-	-	0.13

tiple short windows into a coherent final prediction for longer segments. The method was evaluated on four canines from the AES dataset, achieving a mean sensitivity of 0.84, a time-in-warning of 0.27, and a FPR of 0.78 per day. This approach demonstrated its effectiveness in predicting lead seizures over periods ranging from 169 to 365 days, requiring only two lead seizures for initial model training. The novel handling of sparse and unbalanced data through a group learning approach, coupled with adaptive post-processing, significantly enhances the prediction capability, making the method highly accurate and robust over extended periods. Ra et al. [48] employed PE and a genetic algorithm (GA) combined with kNN to optimize EEG channel selection for more effective feature extraction to differentiate between preictal and interictal states. SVM was then used to classify preictal and interictal states. The results evaluated on CHB-MIT dataset showed an average prediction rate of 92.42% across the 22 patients with optimized channel selection, significantly higher than the 71.13% obtained using all channels. The channel selection method also led to an average increase of 10.58% in accuracy, 23.57% in sensitivity, and 5.56% in specificity. These findings underscore the significant improvement in seizure prediction efficiency by customizing EEG channel selection according to individual patient characteristics. Yang et al. [40] employed PE as the primary feature to analyze iEEG signals on a per-channel basis using a sliding window, capturing the complexity of brain activity in patients with intractable focal epilepsy. Then a SVM equipped with a radial basis function (RBF) kernel was used on patient-specific EEG data from 19 patients on the Freiburg Hospital dataset, with parameters optimized through grid search to differentiate between preictal and interictal states effectively. The method demonstrated an impressive average sensitivity of 94% across patients, with a FPR of only 0.111 per hour. Remarkably, some patients exhibited a sensitivity of 100% with a FPR of zero, highlighting the ability to predict seizure events accurately. Kitano et al. [35] utilized the DWT on 4-second non-overlapping windows through the Haar wavelet function to extract features. The wavelet transforms and corresponding zero-crossing analysis of wavelet detail coefficients were carried out to select features. For classification, self-organizing maps (SOM), as an unsupervised learning algorithm, was used to distinguish between preictal and interictal states. This method achieved high performance metrics, with an sensitivity up to 98%, a specificity up to 88%, and an accuracy up to 91% across the patients studied on the CHB-MIT dataset. These results were derived from a polling-based decision process that classified EEG data into preictal and interictal states, demonstrating the approach's effectiveness in predicting seizures in a clinical setting. Qureshi et al. [49] proposed a computationally efficient method using discrete wavelet decomposition and SVM classifiers on 1-hour EEG recordings from the CHB-MIT dataset. It achieved a sensitivity of 94.9%, an accuracy of 97.43%, and a FPR of 0.138 per hour using just 1-2 channels, making it efficient for real-time and resourceconstrained applications. Costa et al. [50] utilized patient-specific machine learning models, including logistic regression, SVM, and shallow neural networks, to predict seizures using EEG data from 40 patients in EPILEPSIAE dataset. The EEG signals were divided into non-overlapping 5-second windows to extract relevant time-frequency features. Logistic regression demonstrated the best performance, with a seizure sensitivity of 0.13, an FPR of 0.36 per hour, and an improvement over the chance of 12.5%.

2.4.2 Deep Learning

Tsiouris et al. [33] involved time and frequency domain characteristics, cross-correlation between EEG channels, and graph-theoretic features as feature extraction for predicting epileptic seizures. Then, these features were standardized and directly fed into LSTM. Evaluation of seizure prediction performance was conducted using different lengths of preictal windows, ranging from 15 minutes to 2 hours. The results on the CHB-MIT dataset showed that the method achieved high sensitivity and maintained low FPR between 0.11 to 0.02 false alarms per hour. Ultimately, the method exhibited superior performance compared to previous methods evaluated on the same dataset. Truong et al. [34] performed feature extraction using the STFT on 30-second EEG windows to capture both frequency and time domain features from three distinct datasets. Then, the classification was conducted using CNN for the ability to predict seizures. The results were impressive, with the method achieving a sensitivity of 81.4% and a FPR of 0.06 per hour on the Freiburg Hospital dataset, a sensitivity of 81.2% and a FPR of 0.16 per hour on the CHB-MIT dataset, and a sensitivity of 75% with a FPR of 0.21 per hour on the AES dataset. These findings illustrate the robust performance of the method across multiple datasets. Daoud et al. [43] explored four types of deep learning models to effectively predict seizures by feeding raw EEG signals. A channel selection algorithm was utilized to enhance computational efficiency, which is crucial for real-time applications. The results evaluated on the CHB-MIT dataset varied across the different models. The MLPs showed lower performance, with a sensitivity of 84.67% and a specificity of 82.60%. An improvement was noted with the combination of DCNN and MLPs, which achieved a sensitivity of 95.41% and a specificity of 92.80%. The combination of DCNN and Bi-LSTM demonstrated the highest performance, with an impressive sensitivity of 99.72% and a specificity of 99.60%, along with a very low FPR of 0.004 per hour. The combination of DCAE and Bi-LSTM matched the high performance of the combination of DCNN and Bi-LSTM, but offered the advantage of reduced training time due to efficient initial parameter setting facilitated by unsupervised pre-training of the encoder. These findings underscore the potential of end-to-end and integrated deep learning frameworks to improve the accuracy and efficiency of seizure prediction systems. Hussein et al. [37] presented a feature extraction

Table 2.5 Comparative of classification methods by deep learning algorithms

Method	Year	Dataset	Preictal Time (mins)	Number of Subjects	Feature Extraction	Classifiers	Accuracy (%)	Sensitivity (%)	Specificity (%)	FPR (/h)	AUC
Tsiouris et al. [33]	2018	CHB-MIT	15, 30, 60, 120	23 pediatric patients	Time-frequency domain features, cross-correlation, graph features	LSTM	-	99.28, 99.37, 99.63, 99.84	99.28, 99.60, 99.78, 99.86	0.107, 0.063, 0.032, 0.02	-
Truong et al. [34]	2018	Freiburg, CHB-MIT, AES	-	13 patients, 13 patients, 5 dogs and 2 patients	STFT	CNN	-	81.4, 81.2, 75	-	0.06, 0.16, 0.21	-
Daoud et al. [43]	2019	CHB-MIT	60	8 pediatric patients	Raw EEG	Bi-LSTM with DCAE	-	99.72	99.60	0.004	-
Hussein et al. [37]	2021	CHB-MIT, AES, Melbourne	-	23 patients, 5 dogs and 2 petients, 3 patients	CWT	SDCN	99.72, -, -	99.78, 88.45, 89.52	99.60, -, -	0.04, -, -	-, 0.928, 0.883
Truong et al. [39]	2019	CHB-MIT, Freiburg, EPILEPSIAE	-	23 patients, 13 patients, 30 patients	GAN	Two fully-connected layers	-	-	-	-	0.777, 0.755, 0.651
Korshunova et al. [38]	2018	AES	60	5 dogs and 2 patients	Logarithm of the Fourier amplitude	CNN	-	-	-	-	0.81
Zhao et al. [19]	2022	CHB-MIT, AES, Melbourne	30, 60, 60	10 patients, 5 dogs and 2 patients, 3 patients	Raw EEG	CNN	-	99.81, 93.48, 85.19	-	0.005, 0.063, 0.116	1, 0.977, 0.933
Xu et al. [44]	2020	AES, CHB-MIT	60, 30	5 dogs, 7 patients	Raw EEG	CNN	-	93.5, 98.8	-	0.063, 0.074	0.981, 0.988
Li et al. [42]	2023	CHB-MIT, AES	-	18 patients, 4 dogs	Denoising-weighted block	MLPs	-	96.6, 92.9	86.7, 92.1	0.060, 0.025	93.8, 96.5
Wu et al. [45]	2023	CHB-MIT	30	13 patients	Gamma band of raw EEG	LSTM	-	92.17	-	0.27	-
Lee et al. [51]	2024	CHB-MIT	30	24 patients	STFT	Pre-train + ResNet-LSTM	-	81.54	-	0.073	-
Shi et al. [52]	2024	AES, CHB-MIT	60, 30	5 dogs, 13 patients	STFT	B2-ViT Net	-	85.2, 93.3	-	0.013, 0.057	0.816, 0.923

method, converting time-series EEG signals into scalograms through CWT. This allowed the use of semi-dilated convolutional neural networks (SDCN), which handled the unique dimensions of non-square scalogram, to predict seizures. The results on various datasets revealed high performance in seizure prediction, with the method achieving an accuracy of 99.72%, a sensitivity of 99.78%, a specificity of 99.60%, and a FPR of 0.04 per hour on the CHB-MIT dataset, a sensitivity of 88.45% and an AUC of 0.928 on the AES dataset, a sensitivity of 89.52% and an AUC of 0.883 on the Melbourne University dataset. Compared to traditional methods and other deep learning models, the method demonstrated superior performance with consistently low false positive rates and high sensitivity. Korshunova et al. [38] started with feature extraction where EEG signals were resampled to 400 Hz and filtered between 0.1 and 180 Hz, and then partitioned into 1-minute frames for Fourier transformation to extract spectral power in discrete frequency bands. For classification, a CNN was employed to achieve an AUC of 0.81 on the AES dataset. This demonstrated its ability to recognize intricate patterns in EEG signals, surpassing traditional methods such as SVM and linear discriminant analysis (LDA) in performance. Truong et al. [39] introduced an innovative approach using GAN for feature extraction, which is particularly effective in handling unstructured EEG signals. The EEG signals was preprocessed by applying STFT to 28-second windows, preparing the input for the GAN architecture. The discriminator part of the GAN was then used to extract relevant features, classified using two fully-connected layers, designed to identify patterns that suggest upcoming seizures. The method achieved an AUC of 77.68% on the CHB-MIT dataset, 75.47% on the Freiburg Hospital dataset, and 65.05% on the EPILEPSIAE dataset. These outcomes validated the utility of GAN in unsupervised feature extraction and underscored their potential for improving seizure forecasting. Wu et al. [45] involved directly inputting filtered EEG signals, specifically from the gamma band, into the LSTM. This end-to-end method allowed the LSTM to autonomously learn deep features crucial for seizure prediction. The classification strategy was distinct in that it focused on the preictal and interictal stages and incorporated analysis of the postictal stage to enhance overall prediction performance. The method initially achieved a sensitivity of 91.76% with a FPR of 0.29 per hour. The integration of postictal stage analysis further increased the sensitivity to 92.17% and reduced the FPR to 0.27 per hour. Moreover, the method provided a warning time of 44.46 minutes for taking intervention measures, underscoring its potential for timely and effective seizure management. Li et al. [42] presented a novel end-to-end method comprising a denoising-weighted block and a MLPs block for epilepsy seizure prediction. The denoising-weighted block included a denoising layer to eliminate artifacts, a weighted layer to assign importance to each EEG channel, and a reduction layer to decrease the length of the EEG inputs. Subsequently, the MLPs block utilized an inter-channel layer to analyze relations between channels and an intra-channel layer to extract information within each channel. The method achieved a sensitivity of 96.6%, a specificity of 86.7%, an AUC of 0.938, and a FPR of 0.060 per hour on the CHB-MIT dataset. The performance on the AES dataset was also notable, with a sensitivity of 92.9%, a specificity of 92.1%, an AUC of 0.965, and a FPR of 0.025 per hour. The proposed method demonstrated high performance across two datasets and showed the utility of integrating inter-channel and intra-channel information in enhancing prediction outcomes. Xu et al. [44] utilized both one-dimensional and two-dimensional convolutional kernels within the network to predict epileptic seizures from raw EEG signals. The network was designed to capture temporal and spatial dimensions effectively. Initial layers employed one-dimensional kernels to preserve temporal information, while subsequent layers used two-dimensional kernels to integrate spatial information across different channels. The method revealed a sensitivity of 93.5%, a FPR of 0.063 per hour, and an AUC of 0.981 on the AES dataset, a sensitivity of 98.8%, a FPR of 0.074 per hour, and an AUC of 0.988 on the CHB-MIT dataset. These metrics demonstrate the method's robust seizure prediction performance across different types of EEG data. Zhao et al. [19] utilized NAS to automatically design an efficient network for low power consumption while maintaining high performance. To further enhance the model's energy efficiency, techniques such as pruning, quantization, and compact network design were implemented, significantly reducing the size and energy consumption of the model. The results showed a sensitivity of 99.81% , a FPR of 0.005 per hour, and an AUC of 1 on the CHB-MIT dataset. The AES dataset recorded a sensitivity of 93.48%, a FPR of 0.063 per hour, and an AUC of 0.977. The Melbourne University dataset attained a sensitivity of 85.19%, a FPR of 0.116 per hour, and an AUC of 0.933. The performance highlighted the method's potential for real-time application in biomedical devices, providing a sustainable solution for continuous, long-term monitoring and seizure prediction with minimal energy consumption. Lee et al. [51] proposed a hybrid ResNet-LSTM model for epileptic seizure prediction, pre-trained with supervised contrastive learning, achieving an average sensitivity of 81.54%, FPR of 0.073, and accuracy of 87.12% on the CHB-MIT dataset with a 30-minute preictal duration. Shi et al. [52] developed a seizure prediction model based on a Broad Vision Transformer with broad attention (B2-ViT) to capture global spatial interactions and long-range temporal dependencies, achieving AUC, sensitivity, and FPR of 0.923, 93.3%, and 0.057/h on the CHB-MIT dataset, and 0.816, 85.2%, and 0.013/h on the AES dataset, while offering model interpretability through channel attention evaluation.

2.5 Vision and Discussion

The vision for seizure prediction in resource-limited wearable devices includes the development of high-performance, energy-efficient, and interpretable methods that can operate in real time. These methods should be capable of providing early warnings of seizures to improve the quality of life for individuals with epilepsy, especially those with drug-resistant forms of the condition. The ultimate goal is to integrate the methods into compact, low-power wearable devices that are comfortable for daily use, ensuring continuous monitoring and timely intervention with minimal intrusion.

Despite significant efforts to improve the prediction of epileptic seizures, current methods and algorithms have not yet advanced to the point where they can be transformed into commercially viable wearable medical devices. The literature review reveals that traditional ML-based methods for seizure prediction are progressively being replaced by DL-based counterparts. Traditional ML-based methods, unfortunately, yield lower AUC, which serves as a significant metric for assessing its efficacy in alleviating patient anxiety (low false alarm rate) and enabling preemptive medication interventions (high sensitivity). DL-based methods demonstrate higher AUC, indicating superior performance in predicting epilepsy. However, this enhanced performance comes with a notable downside: a substantial increase in energy consumption. This trade-off highlights the ongoing challenge within resource-limited wearable medical devices to balance performance with energy consumption, as originally discussed by Zhao et al. [19], urging further research into optimizing DL-based approaches for real-time epileptic seizure prediction.

Meanwhile, the literature review reveals a notable lack of focus on interpretability for DL-based seizure prediction methods. Despite the advanced capabilities of DL-based methods in handling complex data and improving performance, their black-box nature raises concerns, particularly in the medical field for clinicians. Interpretability is crucial for clinicians to trust and use these methods effectively, as it provides information on the algorithm decision-making process. It highlights the need and efforts to develop DL-based methods that improve performance and make these methods transparent and understandable to healthcare professionals.

CHAPTER 3 THEORY AND METHODOLOGY

This chapter comprehensively introduces the theory and methodology used in the following contribution chapters. Firstly, this chapter presents the general layers and training principles of 1D CNN, which is essential for Chapter 4. Then, it outlines the fundamental principles of quantization in Chapter 5. Lastly, it covers the methodology of the attention mechanism and gradient-weighted class activation mapping (Grad-CAM) in Chapter 6.

3.1 1D CNN

3.1.1 Layers in 1D CNN

A 1D CNN [53] comprises multiple building blocks called layers. Each layer performs specific transformations on its input and passes the result to the next layer. The primary layers include the convolutional layer, the pooling layer, and the fully connected layer, each serving a different function in the network.

Convolutional Layer

The convolutional layer is the core building block of an 1D CNN. It performs a one-dimensional convolution operation that filters the input to extract features. The one-dimensional convolution operates on one-dimensional sequence data, making it computationally efficient for tasks where model deployment efficiency is critical. The output y_j at each position j is calculated as a sum of the product of inputs x and kernel weights w, offset by a bias term b_j. This one-dimensional convolution operation is formally expressed as:

$$y_j = b_j + \sum_{c=0}^{n_c-1} \sum_{k=-p}^{p} x_{c,j-k} w_{c,k} \, , \tag{3.1}$$

where n_c represents the number of input channels, $2p+1$ indicates the size of the kernel.

Pooling Layer

A pooling layer is used to reduce the spatial dimensions of the input for the next convolution layer. It helps reduce computational load, memory usage, and the number of parameters.

There are several types of pooling, but max pooling is the most common, defined as:

$$P(x) = \max_{x \in R} x \,, \tag{3.2}$$

where R represents a particular region over which the maximum is taken.

Fully Connected Layer

Towards the end of the network, a fully connected layer is used, where every input is connected to every output by a learned weight. Typically, this layer is designed to output the classification results based on the features learned by the convolutional and pooling layers. The operation in a fully connected layer can be represented as:

$$y = Wx + b \,, \tag{3.3}$$

where x is the input to the layer, W represents the weight matrix, b is the bias vector, and y is the output vector.

Activation Function

ReLU Following the convolution operation, a rectified linear unit (ReLU) function is applied to introduce nonlinearity into the network, allowing it to learn more complex patterns, defined as:

$$f(x) = \max(0, x) \,. \tag{3.4}$$

The ReLU function helps the network learn faster and perform better by only activating specific neurons, making the network sparse and efficient.

Tanh The hyperbolic tangent (Tanh) function outputs values between -1 and 1. The formula for the Tanh function is given by:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \,. \tag{3.5}$$

The Tanh function is advantageous because it is zero-centered, which helps in centering the data and making the optimization process more efficient during the training of neural networks.

Softmax The softmax function is typically used in the final layer of a classifier, outputting the probabilities of each class. It is defined as:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} , \tag{3.6}$$

where z is a vector of the inputs to the output layer, K is the number of classes, and i indicates the i-th dimension of the output.

Normalization Layer

A Normalization layer, such as batch normalization (BN), is often used in CNN to make training faster and more stable through normalization of the input layer by adjusting and scaling activations. The formula for BN is:

$$\hat{x}^{(k)} = \frac{x^{(k)} - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} , \tag{3.7}$$

where μ_B is the mean and σ_B^2 is the variance of the features in the batch, and ϵ is a small constant to avoid division by zero.

Dropout Layer

Dropout is a regularization method in which randomly selected neurons are ignored during training, reducing the risk of overfitting. This does not directly affect the network architecture, but is critical in the training phase to ensure generalization.

3.1.2 Training Principle

Backpropagation

Backpropagation computes the gradient of the loss function for each weight and bias by the chain rule, propagating the error backward through the network. The input gradient is transmitted back to the preceding layer in the network, continuing the process up the line. An optimizer utilizes the parameter gradient, which consists of the weights gradient and the bias gradient, to modify the parameter values to minimize the loss function.

Input Gradient The gradient of the loss function for each input x is crucial for training, especially when the network includes multiple layers. It is given by:

$$\frac{\partial \mathcal{L}}{\partial x_{c,i}} = \sum_{k=-p}^{p} \frac{\partial \mathcal{L}}{\partial y_{i+k}} w_{c,k} \,. \tag{3.8}$$

Parameter Gradient The gradients for convolutional kernels are computed to update the weights. Each weight gradient is calculated by:

$$\frac{\partial \mathcal{L}}{\partial w_{c,k}} = \sum_{j=0}^{m-1} \frac{\partial \mathcal{L}}{\partial y_j} x_{c,j-k} \,. \tag{3.9}$$

The bias gradient is simpler, directly relating the gradient of the loss to the gradient of the output feature map:

$$\frac{\partial \mathcal{L}}{\partial b_j} = \frac{\partial \mathcal{L}}{\partial y_j} \,. \tag{3.10}$$

Gradient Descent

Adam Adaptive moment estimation (Adam) is an adaptive learning rate optimization algorithm designed to train deep neural networks. It merges the benefits of the most effective aspects of Momentum and root mean square propagation (RMSprop):

$$w^{(l)} = w^{(l)} - \frac{\eta}{\sqrt{\hat{v}} + \epsilon} \cdot \hat{m} \,, \tag{3.11}$$

where $\hat{m}$ and $\hat{v}$ are estimates of the first and second moments of the gradients, respectively, and ϵ is a small scalar used to prevent division by zero.

3.2 Fundamentals of Quantization

Quantization reduces the numerical precision of weights and activations in CNNs, enabling models to be more computationally efficient and less memory-intensive. This section explores uniform quantization methods, quantization granularity, and quantization-aware training [54].

3.2.1 Uniform Quantization Methods

Uniform quantization can be broadly categorized into two types: symmetric and asymmetric. A common approach is to employ symmetric uniform weight quantization and asymmetric uniform activation quantization, which avoids the additional data-dependent term to accelerate computation.

Symmetric Uniform Quantization

In symmetric uniform quantization, values are uniformly quantized such that the intervals between consecutive quantized values are equal. This method does not use a zero-point offset, which simplifies computation but assumes that the distribution of values is centered around zero. The quantization and dequantization processes can be represented as follows:

$$Q(x) = \text{clamp}\left(\left\lfloor \frac{x}{s} \right\rceil ; -2^b + 1, 2^b - 1\right) , \tag{3.12}$$

$$DQ(x) = s \cdot Q(x) \, , \tag{3.13}$$

where x is the original value, $s = \frac{\max(|x_{max}|,|x_{min}|)}{2^b - 1}$ is the scale factor, the clamp$(\cdot)$ function limits all values to the range of $-2^b + 1$ to $2^b - 1$, b represents the number of quantization bits.

Pros 1) Without offset, the amount of calculation can be reduced; 2) The weight values distributed on the positive and negative semi-axes can be fully utilized, resulting in a higher utilization rate; 3) Low-precision multiplication instructions can be directly used to speed up operations; 4) This approach can effectively alleviate the weight distribution problem across different ranges.

Cons 1) The quantization bits may be insufficient for data that is distributed close to zero; 2) The quantization performance will be suboptimal if the data distribution is highly dispersed.

Asymmetric Uniform Quantization

Asymmetric uniform quantization allows for a zero-point, enabling the quantization of values that are not symmetric around zero. This is particularly useful for skewed data distributions or when preserving zero values is important. The corresponding formulas are:

$$Q(x) = \text{clamp}\left(\left\lfloor \frac{x}{s} \right\rceil + z; -2^b, 2^b - 1\right)\,, \tag{3.14}$$

$$DQ(x) = s \cdot (Q(x) - z)\,, \tag{3.15}$$

where z is the zero-point. where x is the original value, $s = \frac{x_{max}-x_{min}}{2^{b+1}-1}$ is the scale factor, $z = 2^b - 1 - \left\lfloor \frac{x_{max}}{s} \right\rceil$ is the offset, the clamp$(\cdot)$ function limits all values to the range of -2^b to $2^b - 1$, b represents the number of quantization bits.

Pros 1) The offset enhances the quantization resolution; 2) This approach is particularly suitable for situations where the data distribution range is relatively concentrated.

Cons 1) The calculation of the offset requires additional storage space, thereby increasing memory usage; 2) Determining the offset involves addition and subtraction operations, which complicates the quantization process; 3) Low-precision multiplication instructions require additional bias operations, which increases the amount of calculation.

3.2.2 Quantization Granularity

Quantization can be applied at different levels of granularity, such as per-tensor or per-channel, each with specific advantages. To ensure low latency and minimal quantization error, per-tensor quantization is chosen for activation values, while per-channel quantization is selected for weights.

Per-Tensor Quantization

Per-tensor quantization uses a single scale factor and zero-point for the entire tensor. This method is less computationally demanding but may lead to higher quantization error if the data variability across the tensor is high.

Per-Channel Quantization

Per-channel quantization applies a separate scale factor and zero-point for each channel in a tensor. This approach can adapt better to variations within each channel, typically yielding higher model accuracy.

3.2.3 Quantization-Aware Training

Quantization-aware training (QAT) involves modifying the network training process to incorporate quantization effects, thereby reducing the discrepancy between the performance of the quantized model and its full-precision counterpart. For asymmetric uniform quantization, the forward pass can be expressed as:

$$\hat{\mathbf{x}} = s \cdot \left(\text{clamp}\left(\left\lfloor \frac{\mathbf{x}}{s} \right\rceil + z; -2^b, 2^b - 1\right) - z\right), \tag{3.16}$$

while the backward pass uses straight-through estimators (STE) to approximate gradients:

$$\begin{aligned}
\frac{\partial \widehat{\mathbf{x}}_i}{\partial \mathbf{x}_i} &= s \cdot \frac{\partial}{\partial \mathbf{x}_i} \text{clamp}\left(\left\lfloor \frac{\mathbf{x}_i}{s} \right\rceil + z; -2^b, 2^b - 1\right) \\
&= \begin{cases} s \cdot \left(\frac{\partial \lfloor \mathbf{x}_i/s \rceil}{\partial (\mathbf{x}_i/s)}\right) \cdot \frac{\partial (\mathbf{x}_i/s)}{\partial \mathbf{x}_i} & \text{if } -2^b \leq \left\lfloor \frac{\mathbf{x}_i}{s} \right\rceil \leq 2^b - 1, \\ s \cdot \frac{\partial (-2^b)}{\partial \mathbf{x}_i} & \text{if } \left\lfloor \frac{\mathbf{x}_i}{s} \right\rceil < -2^b, \\ s \cdot \frac{\partial (2^b - 1)}{\partial \mathbf{x}_i} & \text{if } \left\lfloor \frac{\mathbf{x}_i}{s} \right\rceil > 2^b - 1, \end{cases} \\
&= \begin{cases} 1, & \text{if } -2^b \leq \left\lfloor \frac{\mathbf{x}_i}{s} \right\rceil \leq 2^b - 1, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{3.17}$$

3.3 Attention and Visualization

3.3.1 Attention Mechanism

Attention mechanism [55] allows neural networks to focus on the relevant parts of the input when making predictions. They dynamically weigh the importance of different elements in the input, enabling more effective information processing.

Spectral Attention Module

The spectral attention module aims to enhance the learning of spectral characteristics associated with onset seizures. The basic structure of the spatial attention module is shown in Figure 3.1. Assume that an input feature map Y_{IN} is expressed as $Y_{IN} = [y_1, y_2, \ldots, y_F]$, where each $y_i \in \mathbb{R}^{1 \times C}$ and $i = 1, 2, \ldots, F$. First, $Conv1$ operation gets spectral information across channels through a kernel size of 1. After the Sigmoid function, the spectral recalibration vector m can be described as:

$$m = \sigma(Conv1(Y_{IN})), \tag{3.18}$$

where σ is the Sigmoid function. $Conv1$ aggregates information across channels to identify significant features associated with frequencies, so m_i indicates the importance of the ith frequency range. Then, m is used to recalibrate the input feature map Y_{IN} to

$$Y_c = m \cdot \delta(Conv2(Y_{IN}))\,, \tag{3.19}$$

where δ is the ReLU function and $Conv2$ represents a convolutional operation that encodes feature information across channels to avoid excessive emphasis on different channels. Since m is between 0 and 1, stacked recalibration of features can reduce the values of deep features, which may lead to vanishing gradients. To mitigate this issue, a residual connection [56] is employed. As a result, the final output is $Y_{OUT} = Y_{IN} + Y_c$.

Spatial Attention Module

The spatial attention module improves network performance by adaptively recalibrating channel-wise features in a feature map. The basic structure of the spatial attention module is shown in Figure 3.2. Given an input feature map $Y_{IN} = [y_1, y_2, \ldots, y_C]$, where $y_j \in \mathbb{R}^{F\times 1}$. First, a global average pooling compresses the frequency information to produce a channel-wise vector z, where $z \in \mathbb{R}^{1\times C}$. The jth element of z can be expressed as:

$$z_j = \frac{1}{1\times F}\sum_{k=1}^{F} y_j(k)\,. \tag{3.20}$$

Then z is transformed through two fully connected layers to capture channel-wise dependencies and generate the spatial recalibration vector r, described by:

$$r = \sigma(Conv4(\delta(Conv3(z))))\,, \tag{3.21}$$

where δ is the ReLU function, σ is the Sigmoid function. $Conv3$ and $Conv4$ denote convolutional operations to capture dependencies among channels, each utilizing a kernel size of 1. Next, r is used to recalibrate the input feature map Y_{IN} to

$$Y_c = r \cdot Y_{IN} = [y_1 r_1, y_2 r_2, \ldots, y_C r_C]\,, \tag{3.22}$$

where r_j represents the significance of the jth channel. Thus, the obtained Y_c fully incorporates global spatial information. Since r is between 0 and 1, similar to the spectral attention module, a residual connection [56] is also used to prevent the reduction of the feature response value. As a result, the final output is $Y_{OUT} = Y_{IN} + Y_c$.

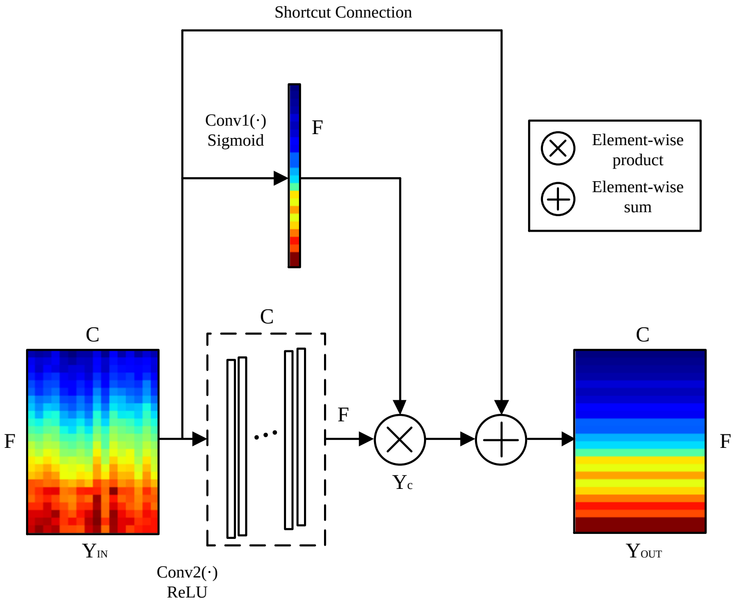


Figure 3.1 The basic structure of the spectral attention module

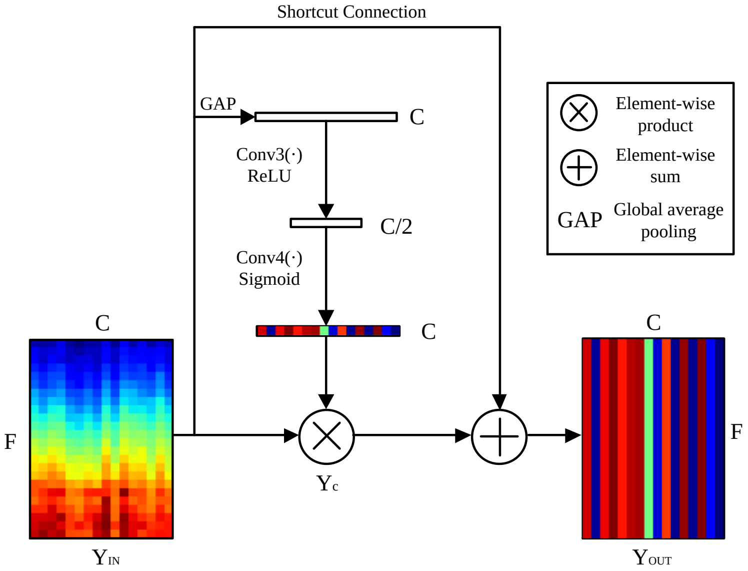


Figure 3.2 The basic structure of the spatial attention module

3.3.2 Grad-CAM

Grad-CAM [57] is a technique used to increase the transparency and interpretability of CNNs by highlighting the important regions in the input for predicting the conceptual output of interest. This is particularly useful for applications where understanding the model reasoning is crucial, such as in medical imaging.

To understand how Grad-CAM works, let Y^c be the score for class c before the softmax layer, and let A^k represent the feature maps at the last convolutional layer. The importance of feature map k for a target class c is given by:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \, , \tag{3.23}$$

where α_k^c denotes the global average pooled gradients. After computing α_k^c, the coarse Grad-CAM localization map $L_{\text{Grad-CAM}}^c$ (of width U and height V) for any class c can be generated as follows:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \, . \tag{3.24}$$

In this expression, the ReLU is applied to the linear combination of maps only to consider the features that positively influence the class of interest, as negative values would represent features that belong to other categories in the scene. The heatmap generated by $L_{\text{Grad-CAM}}^c$ is then overlaid on the original image to show the discriminative regions used by the CNN to identify that class.

3.3.3 Relationship with Interpretability

Attention mechanism and Grad-CAM share the common goal of identifying important parts of the input that contribute to the model's predictions. Attention mechanisms are part of the model architecture and are trained to focus on relevant input parts. Grad-CAM, on the other hand, is a post-training technique that uses gradient information to highlight essential regions. Integrating the attention mechanism with Grad-CAM can provide a more comprehensive understanding of model behavior. Both methods offer interpretability, but the attention mechanism can do so inherently during the prediction process, whereas Grad-CAM provides a visual explanation after the prediction is made.

CHAPTER 4 ARTICLE 1: TINY CNN FOR SEIZURE PREDICTION IN WEARABLE BIOMEDICAL DEVICES

Yang Zhang[1], Yvon Savaria[1], Shiqi Zhao[2], Gonçalo Mordido[1,3],
Mohamad Sawan[1,2], François Leduc-Primeau[1]
[1]Department of Electrical Engineering, Polytechnique Montreal, Canada
[2]School of Engineering, Westlake University, Hangzhou, Zhejiang, China
[3]Mila - Quebec AI Institute, Montreal, Canada

Published in: the 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Publication date: September 8, 2022

The first objective is to design a high-performance, patient-specific CNN with a tiny model size for effective seizure prediction, which helps alleviate patient anxiety (low FPR) and enable preventive measures or medication interventions (high sensitivity). To achieve the first objective, a 1DSCNN is proposed to predict epileptic seizures. This model features a highly competitive, tiny size ideal for integrating into wearable biomedical devices. Compared to state-of-the-art methods, the proposed 1DSCNN reaches top performance with the smallest model size when tested on the AES dataset. Furthermore, a preliminary quantization approach is employed to assess how different bit widths affect model performance, thus aiding its implementation in wearable medical devices.

4.1 Abstract

Epilepsy is a life-threatening disease affecting millions of people all over the world. Artificial intelligence epileptic predictors offer excellent potential to improve epilepsy therapy. Particularly, deep learning models such as CNN can be used to accurately detect ictogenesis through deep structured learning representations. In this work, a tiny 1DSCNN is proposed based on STFT to predict epileptic seizure. The results demonstrate that the proposed method obtains better performance compared to recent state-of-the-art methods, achieving an average sensitivity of 94.44%, average FPR of 0.011/h and average AUC of 0.979 on the test set of the AES dataset, while featuring a model size of only 21.32 kB. Furthermore, after adapting the model to 4-bit quantization, its size is significantly decreased by 7.08x with only 0.51% AUC score precision loss, which shows excellent potential for hardware-friendly wearable implementation.

4.2 Introduction

Nearly 60 million people in the world suffer from epilepsy, a common and serious brain disease which can affect people of all ages [16]. Epilepsy is characterized by unprovoked seizures, and can cause other health problems [58], which may be life-threatening. Long-time medication is a common method to control epilepsy, which can cause some undesirable side effects such as medication resistance [38]. It is especially important for epileptic patients to know when a seizure will happen to allow taking suitable mitigation measures in advance. To this end, seizure prediction can help them improve their well-being.

In clinical practice, brain electrical activities can be measured by multiple channels through a collection of electrodes installed on the scalp or exposed surface of the brain to collect sEEG signals and intracranial EEG signals, respectively [59]. Thus long-term EEG monitoring to process neural signals is critical to patients due to the chronic characteristic of epilepsy. Hence, low-power acquisition and processing of EEG signals should be considered for wearable biomedical devices as well as implanted devices.

Seizure prediction is usually viewed as a binary classification problem between the preictal and non-preictal classes [38]. As shown in Figure 4.1, the preictal state is the period before a seizure onset, while a non-preictal state can be one of three states: interictal (seizure-free), ictal (during a seizure) and postictal (after seizure). The main challenge of the seizure prediction problems is classifying signals into preictal and interictal states [38] with high sensitivity and low false alarm rate, which validates prediction while minimizing disturbance to the normal patients activities [60]. To this end, the recent development of promising deep learning techniques enabled significant performance improvements of seizure prediction methods. Truong et al. [34] developed a generalized retrospective and patient-specific seizure prediction method using STFT and CNNs, with average sensitivity of 75% and average FPR of 0.21/h on the American Epilepsy Society seizure prediction challenge dataset. Zhao et al. [19] explored energy-efficient seizure prediction, through feeding a direct end-to-end time-domain signal to a CNN, with an average sensitivity of 93.48%, average FPR of 0.063/h, and average AUC of 0.977 on the same dataset. Liu et al. [61] proposed a multi-view CNN to predict seizures, by combining time domain and frequency domain features in a CNN, with an average AUC of 0.837 on the same dataset.

Although, deep learning methods often achieve state-of-the-art results compared to traditional machine learning approaches, power consumption is significantly increased due to the size of existing deep learning models. By analyzing the trade-off between power consumption and performance, we propose means to reduce energy consumption while minimizing perfor-

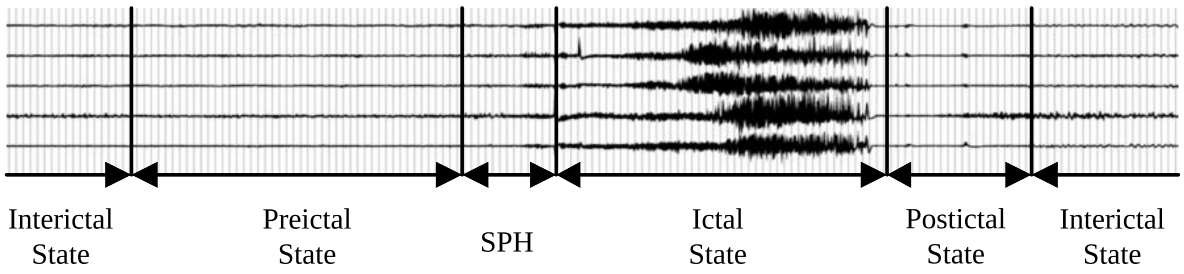


Figure 4.1 An example of multi-channel EEG recording for canine. Interictal state means seizure-free period, whereas ictal relates to the seizure onset period. The preictal state is the pre-seizure period, while the postictal state is the post-seizure period. The SPH represents a period between preictal state and ictal state, where it is ideal to execute medical intervention or to apply risk mitigation measures

mance losses. To achieve this, a hardware-friendly tiny CNN for epileptic seizure prediction is proposed. Our main contributions can be summarized as follows:

- A 1DSCNN is proposed to predict epilepsy seizure for wearable biomedical devices. The proposed method outperforms existing methods in spite of a very competitive small model size.
- Various quantization schemes are applied to the proposed 1DSCNN model for evaluating the impact of different bit widths on model performance.

The remainder of this paper includes the description of the adopted methodology in Section 4.3, our results are reported and discussed in Section 4.4, finally, our main findings and conclusions are reported in Section 4.5.

4.3 Methodology

4.3.1 Dataset

In this work, the challenging and widely used dataset provided during the American Epilepsy Society Seizure Prediction Challenge [62] is adopted as the benchmark dataset to compare various prediction methods. The dataset consists of iEEG recordings from five dogs and two human patients with naturally occurring epilepsy. Details about the collected information using an ambulatory monitoring system are shown in Table 4.1. More specifically, recordings from four of the dogs are obtained through 16 subdural electrodes and the remaining dog through 15 electrodes, sampled at 400 Hz. One patient is recorded through 15 subdural

Table 4.1 Per-subject characteristics of the dataset: number of channels, size of the preictal segments, size of the interictal segments, and interictal hours

Subject	Channels	Preictal segments	Interictal segments	Interictal hours
Dog 1	16	24	480	80.0
Dog 2	16	42	500	83.3
Dog 3	16	72	1440	240.0
Dog 4	16	97	804	134.0
Dog 5	15	30	450	75.0
Patient 1	15	18	50	8.3
Patient 2	24	18	42	7.0

electrodes, while the other patient through 24 electrodes, sampled at 5000 Hz. These are long-duration recordings, spanning from multiple months and up to a year. Also, in this dataset, interictal data segments are required to be at least one week before or after any seizure, while preictal data segments cover one hour before a seizure with a five-minute seizure horizon. The annotated dataset is divided into two parts for each subject through five-fold stratified cross-validation [63]: 80% training set and 20% testing set. Then, 20% of the training set is used as validation set.

4.3.2 Preprocessing

The preprocessing stage consists of data segmentation, resampling, and a STFT. During data segmentation, each 10-minute preictal or interictal segment is sliced into 20-second clips without overlap to augment the dataset. Then, each 20-second clip is resampled at 400 Hz for convenient processing. Before feeding the data to CNN, the initial raw iEEG data is converted into a two-dimensional time-frequency representation. Based on the non-stationary nature of iEEG signals, which highly depend on time, STFT is employed to convert time-series EEG signals into time-varying frequency components [64].

The raw iEEG signal is converted into 24 frequency bands according to brain wave frequencies from 0.1 Hz to 190 Hz as shown in Figure 4.2. More specifically, delta (0.1-2, 2-4 Hz), theta (4-6, 6-8 Hz), alpha (8-10, 10-12 Hz), beta (12-21, 21-30 Hz), low-gamma (30-40, 40-50, 50-60, 60-70, 70-80, 80-90, 90-100 Hz) and high-gamma (100-110, 110-120, 120-130, 130-140, 140-150, 150-160, 160-170, 170-180, 180-190 Hz) [65]. When applying STFT, there is a trade-off between time resolution and frequency resolution. The 20-second window length

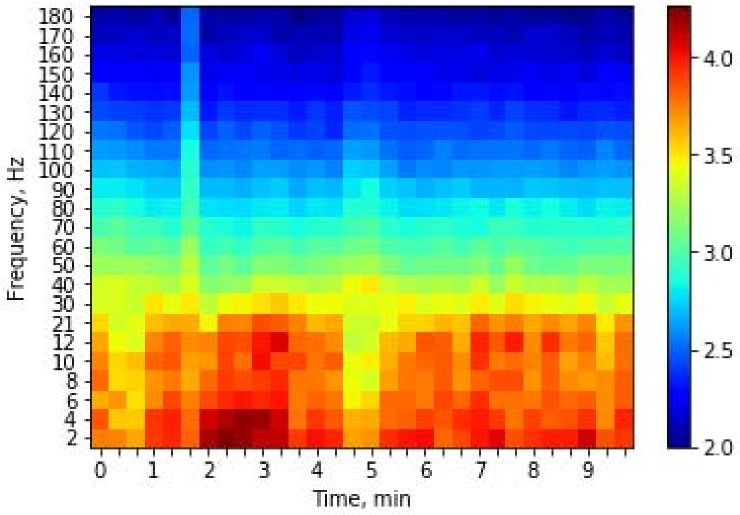


Figure 4.2 The mean value of spectrum amplitude in 24 frequency bands from 0.1 to 190 Hz of a 10-minute segment for a single channel

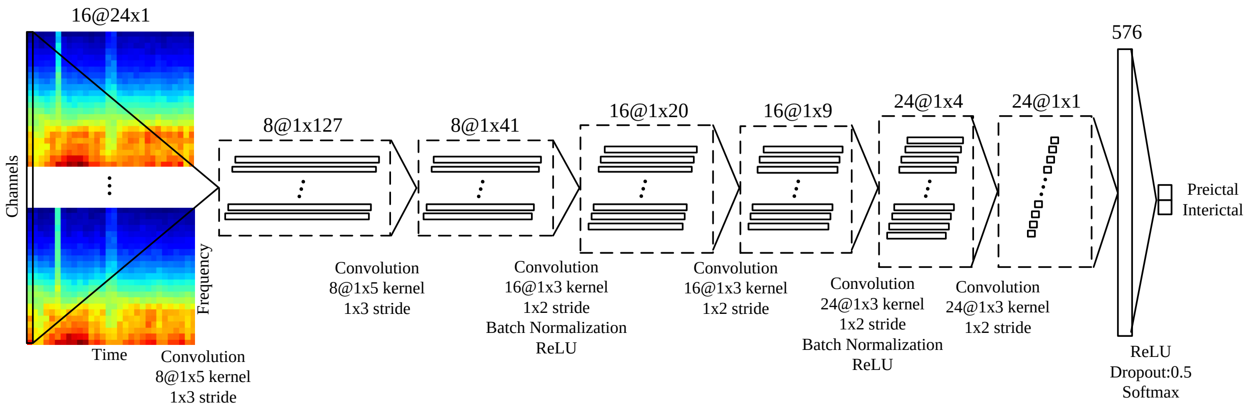


Figure 4.3 Architecture of the proposed convolutional neural network

is selected to guarantee a frequency resolution over 0.1 Hz. A rectangular window shape is used because it reduces the main lobe width in the frequency domain, thus improving the frequency resolution [66]. Given an occurrence of abnormal brain discharge, the energy of the pre-seizure state is assumed to be concentrated in certain frequency bands. Hence, to reduce complexity and foster deployment on wearable biomedical devices, the mean value of the spectrum amplitude in each band is proposed as input features for the CNN. For each subject, the input consists of a 20-second clip, the CNN outputs one prediction for every time clip. The input size is $Number\ of\ channels \times 24$.

4.3.3 CNN Architecture and Training Settings

With the recent developments in bioinformatics, CNN is an attractive approach to analyze EEG signals [67] through the extraction of low-level features to be fed in subsequent layers to represent high-level features. In this work, a 1DSCNN is proposed to predict epilepsy seizure. The overall CNN architecture is shown in Figure 4.3. The stacked convolutional layer, initially proposed in VGGNet [68], presents two advantages: the depth of the neural network is improved and the amount of parameters is reduced under the condition of ensuring the same receptive field. Compared to two-dimensional CNN (2D CNN), 1D CNN can extract not only interior image pixels, but also more details about low-level features, such as edge shape, among multi-channels. As described in Figure 4.3, firstly, 16-channel iEEG signals are passed through one 1DSCNN block to extract cross information between different channels at the same time. Then, two 1DSCNN blocks follow to improve the generalization of the model. The ReLU function is used for each layer. Finally, a Softmax layer follows to perform classification. The Adam optimizer is used for training, with a varying learning rate from 10^{-3} to 5^{-4}, and of 0.9 and 0.999 respectively. The learning rate is decreased if the validation error is not improved. Batch normalization and dropout are applied during training to prevent overfitting. The model in this work is implemented in Python 3.6 using Keras 2.3.1 with a Tensorflow 1.13.1 backend. The model is configured to run in parallel on two NVIDIA Tesla V100 graphics cards.

4.3.4 CNN Quantization

For meeting the time and energy constraints in wearable biomedical devices, we quantize the CNN weights and activations through re-training to reduce computation time, memory requirements, and power consumption [69]. We evaluate the impact of quantization on model performance using different bit widths. During the forward pass, weights and activations are quantized as fixed-point values with the same precision through uniform symmetric signed and uniform asymmetric signed quantization, respectively [70]. The scaling factor is a power of two, which allows the scaling to be computed using bit shifts instead of multipliers [71]. Moreover, the Tanh function is used instead of the ReLU function [19] due to the improvement of AUC scores. During the backward pass, the gradient is propagated by full-precision weights and STE [72]. We note that no quantization is applied to the input and output layers.

4.4 Results and Discussion

4.4.1 Evaluation Metrics

A rigorous evaluation methodology is used to assess model performance on each subject of the dataset through five-fold cross-validation. Sensitivity, FPR and AUC are computed to evaluate our approach and compared with recent state-of-the-art works. Sensitivity is the percentage of correctly classified 20-second seizure clips among the total number of 20-second seizure clips. FPR is defined as the false positive rate per hour [34]. AUC is the area under the ROC curve, which illustrates the diagnostic ability of a given classifier.

4.4.2 Performance Analysis

Table 4.2 shows the evaluation results of the proposed 1DSCNN model, achieving an average sensitivity of 94.44%, an average FPR of 0.011/h, and an average AUC of 0.979 for all subjects on the dataset. Median and deviation values for AUC are presented in Figure 4.4 through five-fold cross-validation, where the yellow line in the box and the edge of the box refer to median and quartile values of AUC, respectively. And the bar of box varies from minimum to maximum values of AUC. It is observed that Dog 1 is the subject for which seizures are hardest to predict, because even after tuning hyperparameters as best as we can, the AUC of Dog 1 remains the lowest among all subjects.

Table 4.3 compares our method with other state-of-the-art methods on the same dataset. Model size is reported for 32-bit full precision (FP) parameters. The comparison results demonstrate the proposed 1DSCNN model achieves the best sensitivity, FPR, and AUC at the lowest model size. It is worth noting that, although the average AUC score of our method is 0.002 higher than Zhao et al. [19], our model has less than half the size of their model, showcasing the potential of our model for wearable biomedical devices.

Figure 4.5 demonstrates AUC scores and model size after different quantization levels. Compared to the FP baseline, 8-bit quantization reduces model size by 3.79 times, with only 0.31% AUC score loss. Moreover, 4-bit quantization reduces model size by 7.08 times with

Table 4.2 Per-subject evaluation results: sensitivity, FPR and AUC

Subject	**Sensitivity(%)**	**FPR(/h)**	**AUC**
Dog 1	91.11	0.013	0.926
Dog 2	97.70	0.001	0.998
Dog 3	95.42	0.003	0.978
Dog 4	92.27	0.003	0.974
Dog 5	96.78	0.001	0.999
Patient 1	97.22	0.008	0.998
Patient 2	90.56	0.045	0.979
Average	94.44	0.011	0.979

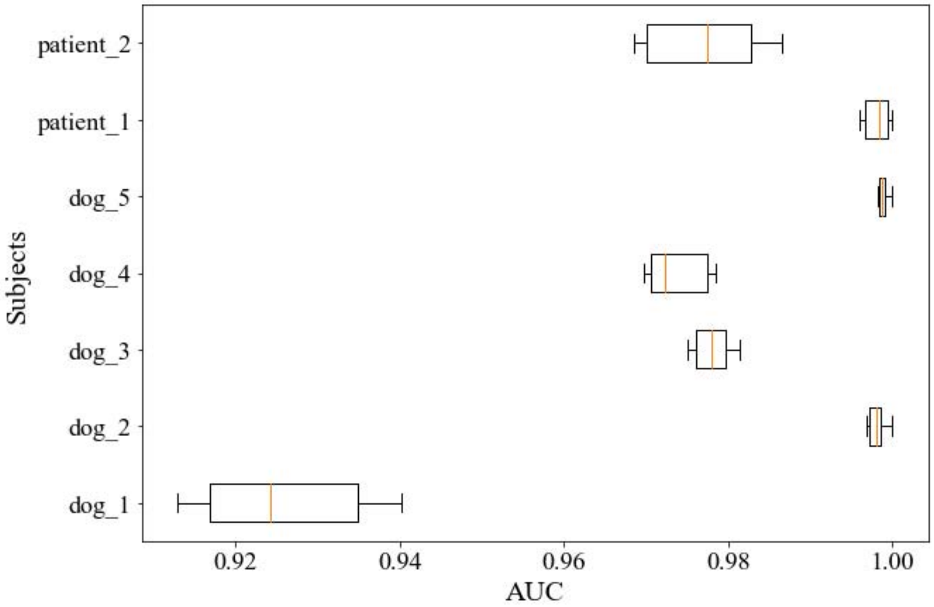


Figure 4.4 Per-subject AUC box plot: median and deviation of AUC scores through five-fold cross-validation

Table 4.3 Comparsion with other state-of-the-art methods

Method	Sensitivity (%)	FPR (/h)	AUC	Model size	
Truong et al. [34]	75.0	0.210	-	0.76 MB	
Zhao et al. [19]	93.48	0.063	0.977	45.22 kB	
Brinkmann et al. [65]	-	-	0.860	-	-
Korshunova et al. [38]	-	-	0.810	0.56 MB	
Liu et al. [61]	-	-	0.837	1.79 MB	
This work	**94.44**	**0.011**	**0.979**	**21.32 kB**	

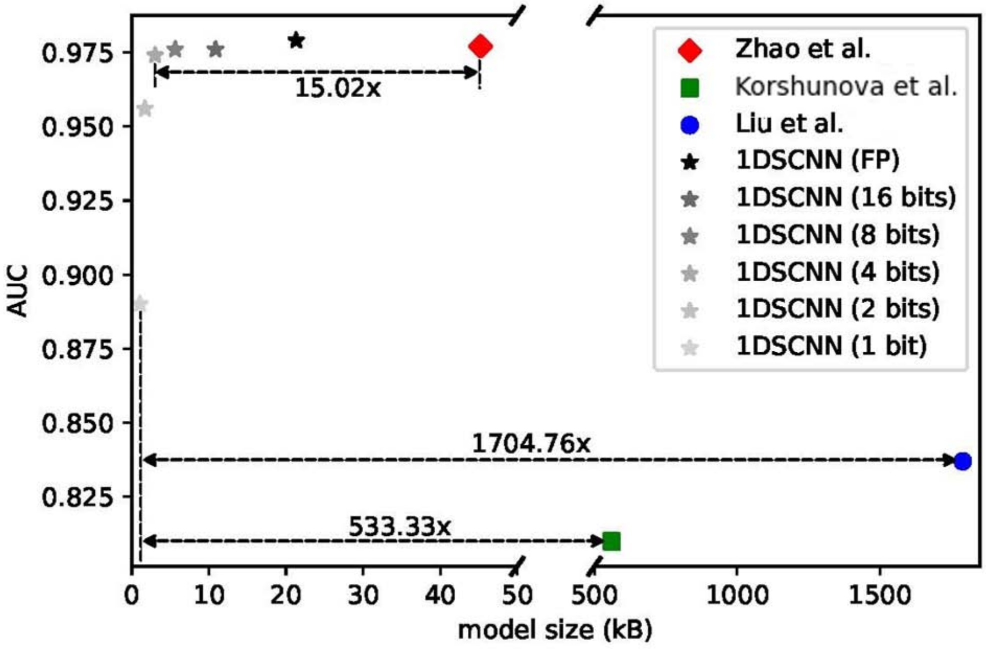


Figure 4.5 Comparison of the proposed methods with other state-of-the-art methods: AUC versus model size

only a 0.51% reduction in AUC score. Finally, 2-bit and 1-bit quantization reduce the model size by 12.54 and 20.30 times with 2.35% and 9.09% AUC score loss, respectively. Hence, 4-bit quantization shows great promise for wearable biomedical devices, significantly reducing the model size at a tolerable precision loss. Figure 4.5 compares results obtained with 1DSCNN implemented at various precisions ranging from floating point to 1 bit with other state-of-the-art methods. Compared to Zhao et al. [19], a 4-bit quantized 1DSCNN reduces the model size 15.02 times with only 0.31% AUC score loss. More notably, compared to Liu et al. [61] and Korshunova et al. [38], 1-bit quantization significantly reduces by 1704.76 and 533.33 times the model size while offering 6.33% and 9.88% AUC score improvement, respectively. The reason why the model size of Liu et al. [61] is so large is that their model exploits two-dimensional inputs and employs more layers. While Korshunova et al. [38] adopts large receptive field filters and input size. These comparisons show the proposed methods outperform existing state-of-the-art methods.

4.5 Conclusion

This paper proposes a 1DSCNN for epilepsy seizure prediction with a model size suitable for wearable biomedical devices. Compared to recent state-of-the-art methods, the proposed 1DSCNN achieves the best performance with the lowest model size on the AES dataset. When combined with quantization, our method is hardware-friendly, easing its deployment in wearable biomedical devices. Further work will consider advanced binary quantization methods of CNN to further improve the AUC of tiny models which are suitable for biomedical implanted devices.

CHAPTER 5 ARTICLE 2: TINY NEURAL NETWORK FOR EPILEPTIC SEIZURE FORECASTING IN WEARABLE DEVICES

Yang Zhang[1], Yvon Savaria[1], Mohamad Sawan[1,2], François Leduc-Primeau[1]
[1]Department of Electrical Engineering, Polytechnique Montreal, Canada
[2]School of Engineering, Westlake University, Hangzhou, Zhejiang, China
Published in: IEEE Transactions on Biomedical Engineering
Submission date: March 20, 2024

The second objective is to explore an energy-efficient, patient-specific DL algorithm for accurately predicting epileptic seizures, addressing the trade-off between performance and energy consumption in resource-limited wearable medical devices. To achieve the second objective, our previously proposed 1DSCNN developed as part of the first objective is further evaluated to show its ability to generalize on the AES and CHB-MIT datasets. In contrast to the AES dataset, the CHB-MIT dataset is primarily used for seizure detection, which requires extra data segmentation and annotation to fulfill the requirements of seizure prediction. Moreover, since the CHB-MIT dataset consists of sEEG recordings that have a lower signal-to-noise ratio compared to the iEEG in the AES dataset, noise suppression is necessary during the preprocessing stage. In contrast to Chapter 4, which only assesses model size, this chapter presents a comprehensive on-chip energy model to evaluate energy costs. Reported results are better than those obtained with previously reported methods regarding sensitivity, FPR, AUC, model size, and energy consumption. This paper also corrects a mistake in the previous description of 1DSCNN in Figure 4.3. Firstly, a multichannel spectrum of EEG signals is processed through a single 1D CNN block to extract cross-channel information, followed by another 1D CNN block to form a 1DSCNN block. Then, two more 1DSCNN blocks are used to enhance generalization. Finally, flatten and dense layers lead to a Softmax layer for classification. This mistake is corrected in Figure 5.3. Subsequently, a fixed-precision quantization method well-suited for energy-limited hardware has been applied to the two datasets, achieving excellent energy efficiency with only minor reductions in performance. This paper presents a more comprehensive quantization method. It shows better outcomes than those reported in Chapter 4 due to various refined elements such as scaling factor, quantization granularity, and training strategy. Finally, to improve performance for hard-to-predict seizures like dog_1 in Chapter 4, a near-optimal, low-complexity mixed-precision search method is also proposed, which can enhance model performance for individuals with such seizures.

5.1 Abstract

Objective: Seizure prediction has emerged as a vital area of research to enhance the lives of drug-resistant epilepsy patients by alleviating their anxiety and facilitating precautionary measures. Many studies have demonstrated that deep learning offers significant advantages in analyzing spatio-temporal non-stationary EEG signals. Transformer-based models cannot generalize well for patient-specific seizure prediction tasks with limited data compared to CNN models. In addition, deep models require large memory footprints and computational costs, which can be impractical for low-power wearables. *Methods:* In this work, a tiny 1DSCNN based on average-pooling STFT is proposed to effectively predict epileptic seizures with low energy consumption. Subsequently, the baseline method has been refined with hardware-friendly fixed-precision and mixed-precision quantization to better serve low-power wearables. *Results:* The baseline model achieves sensitivities of 94.44%, 96.14%, FPR of 0.011/h, 0.018/h, and AUC of 0.979, 0.996, for model sizes of 21.32 kB, 23.62 kB, with estimated energy consumption of 0.20 μJ, 0.28 μJ when processing data from the AES dataset and the CHB-MIT dataset, respectively. Furthermore, the adaptation of the model to 4-bit quantization and mixed-precision quantization results in a model size reduction of 7.08x, 7.06x, with only 0.51%, 0.20% AUC score loss in the AES dataset, and a size reduction of 5.95x, 8.59x, with only 0.20%, 0.30% AUC score loss in the CHB-MIT dataset. *Conclusion/Significance:* These results outperform prior state-of-the-art methods, which shows outstanding potential for hardware-friendly biomedical wearable devices.

5.2 Introduction

Nearly 50 million people in the world suffer from epilepsy, a chronic and lifelong brain disease that can influence people of all ages [16]. Epilepsy is a neurological disease characterized by excessive electrical discharges in some parts of brain cells, which results in life-threatening recurrent seizures [7]. Fortunately, 70% of epileptic patients can enjoy everyday lives with the proper treatment using inexpensive and effective anti-epileptic medicines [7]. However, long-term medication treatment can cause a series of side effects, and medication resistance is one of the worst cases. Thus, it is especially significant for epileptic patients to know when a seizure will occur to allow getting advisable mitigation measures in advance. To this end, seizure prediction can help them recover their well-being.

During a seizure, the electrical activity in the brain is caused by some complex chemical changes in nerve cells. EEG is widely applied in clinical practices to monitor, detect and diagnose epileptic seizures. Those recordings of brain activity are very significant for experts to give a definitive diagnosis of epilepsy. More specifically, brain electrical activities can be accessed by multiple channels through an array of electrodes installed on the scalp or exposed surface of the brain to obtain sEEG signals and iEEG signals, respectively [73]. Hence, it is crucial to research low-power acquisition and processing units to provide real-time long-term prediction of epileptic seizures in wearable devices for patients subject to that chronic disease.

According to different brain discharge patterns, as shown in Figure 5.1, seizure prediction is usually viewed as a binary classification problem to distinguish preictal and non-preictal classes [38]. The preictal (pre-seizure) state is the period before a seizure onset, while the postictal (post-seizure) state is the period after that. The interictal (seizure-free) state means there is no seizure occurring, whereas the ictal (seizure onset) is associated with the occurrence of a seizure. The SPH provides a period to take medical intervention or mitigation measures between the preictal and ictal states. Especially the toughest task for seizure prediction is to divide signals into preictal and interictal states with a high true prediction rate (sensitivity) and low false prediction rate (FPR), which can maximize prediction accuracy while minimizing interruptions to epileptic patients [60].

As mentioned before, long-term brainwave monitoring with wearable devices to predict epileptic seizures could be a precious solution to help sensitive patients improve their well-being. Although seizures can be effectively predicted through advanced deep-learning techniques [37,43,74], the increase in energy consumption with model size cannot be ignored [19,75,76]. Deep learning-based classifiers typically achieve area under the ROC curve (AUC) above 0.95 but demand model sizes exceeding 50 kB and energy consumption surpassing 10 mJ per inference for wearable devices [19]. For this reason, several strategies were proposed to reduce energy consumption while minimizing performance loss through leveraging the trade-off between energy and performance. To this end, a hardware-friendly tiny neural network for epileptic seizure prediction with wearable devices is proposed.

The main contributions of this paper are as follows:

- A computation-efficient feature extraction approach is proposed, which maps spatio-temporal non-stationary EEG signals into pixel-level representations through average-pooling STFT.
- A tiny 1DSCNN baseline model is proposed to effectively predict epileptic seizures for wearable devices, surpassing previous state-of-the-art methods, notably with its compact model size.
- An evaluation of a hardware-friendly fixed-precision quantization scheme has been conducted on the baseline model, with excellent energy efficiency and very low classification

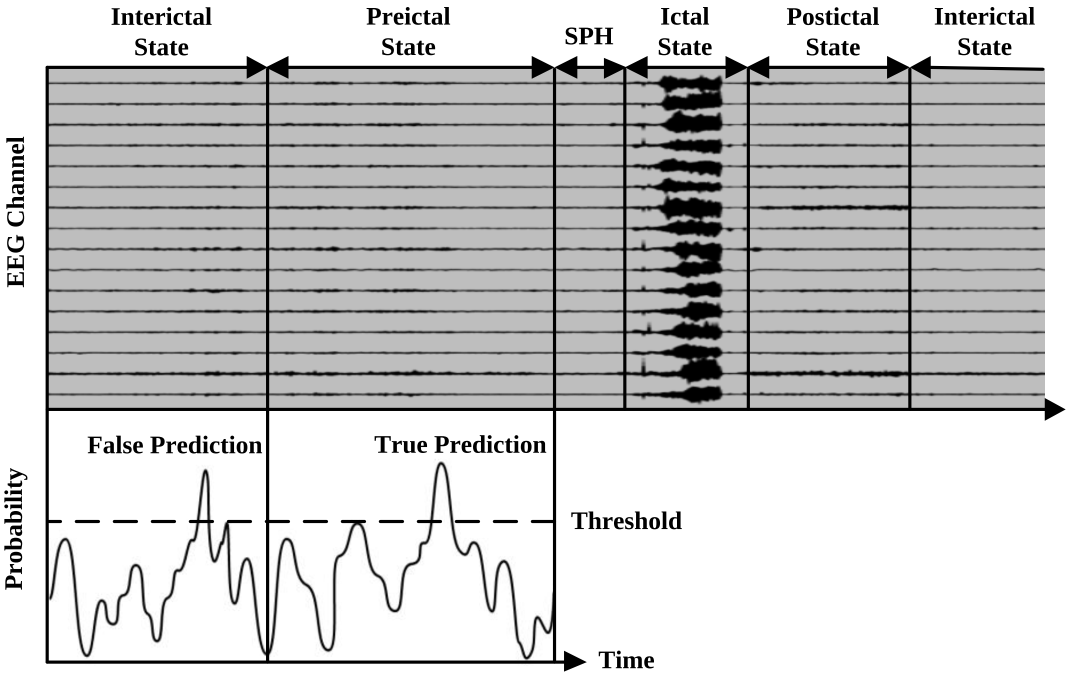


Figure 5.1 Brain discharge patterns of multi-channel EEG recordings

performance losses.

- A near-optimal low-complexity mixed-precision search procedure is further proposed, which can potentially boost the model performance of subjects that have hard-to-predict epileptic seizures.

The remainder of this paper introduces related works, benchmark datasets and evaluation methods in Section 5.3. The details of the adopted methodology are presented in Section 5.4. The results that we obtained are reported, discussed, and compared with related work in Section 5.5, finally, conclusions and the main findings of this work are provided in Section 5.6.

5.3 Background

5.3.1 Related Works

Seizure prediction performance has been notably boosted with the recent development of promising deep-learning techniques. Korshunova et al. [38] proposed a CNN that feeds frequency features with an average AUC of 0.810 on the AES dataset. Liu et al. [61] proposed a multi-view CNN to predict seizures by combining time domain and frequency domain features in a CNN, with an average AUC of 0.837 on the AES dataset. Daoud et al. [43] introduced a method of a front-end deep CNN (DCNN) or a front-end deep convolutional autoencoder (DCAE) with a back-end Bidirectional-LSTM (Bi-LSTM) using raw EEG signals, reaching the same sensitivity of 99.72%, a specificity of 99.60% and a false alarm rate of 0.004/h on the CHB-MIT dataset. Zhang et al. [77] designed a feature extractor combining wavelet packet decomposition, common spatial pattern (CSP), and a shallow CNN, with a sensitivity of 92.2% and FPR of 0.12/h on the CHB-MIT sEEG dataset. Tang et al. [78] developed a multi-view convolutional gated recurrent network (Mv-CGRN), embedded with an attention mechanism that can adaptively tune weight parameters, with an average sensitivity of 94.50% and an average false positive rate of 0.118/h on the CHB-MIT dataset. Wu et al. [75] proposed C^2SP-Net, a neural network designed to jointly address compression, prediction, and reconstruction tasks for epilepsy seizure prediction. Their model achieved a prediction accuracy of 92.5%, a sensitivity of 94.2%, and an FPR of 0.09/h on the CHB-MIT dataset, with an average prediction accuracy loss of only 0.6% across compression ratios ranging from 1/2 to 1/16. Lee et al. [51] proposed a hybrid ResNet-LSTM model for epileptic seizure prediction, pre-trained with supervised contrastive learning, achieving an average sensitivity of 81.54%, FPR of 0.073, and accuracy of 87.12% on the CHB-MIT dataset with a 30-minute preictal duration. Truong et al. [34] developed a generalized retrospective and patient-specific seizure prediction method using STFT and CNNs, with average sensitivity of 75%, 81.2%, and average FPR of 0.21/h, 0.16/h on the AES dataset and the CHB-MIT dataset. Zhao et al. [19] explored energy-efficient seizure prediction by feeding a direct end-to-end time-domain signal to a CNN, with average sensitivity, FPR and AUC of 93.48%, 0.063/h, 0.977 and 99.81%, 0.063/h and 1 on the same datasets. Liang et al. [79] introduced a semi-supervised domain-adaptive seizure prediction method that leverages feature alignment and consistency regularization for cross-patient generalization, achieving average sensitivity, FPR, and AUC of 88.8%, 0.182/h, and 0.849 on the CHB-MIT dataset, and 75.7%, 0.165/h, and 0.763 on the AES dataset. Shi et al. [52] developed a seizure prediction model based on a Broad Vision Transformer with broad attention (B2-ViT) to capture global spatial interactions and long-range temporal dependencies, achieving AUC, sensitivity, and FPR of 0.923, 93.3%, and 0.057/h on the CHB-MIT dataset, and 0.816, 85.2%, and 0.013/h on the AES dataset, while offering model interpretability through channel attention evaluation.

Despite the significant achievements of deep-learning methods, there are evident limitations when considering hardware implementation for wearable devices. These limitations include overlooked model parameters [34, 37, 51, 61, 77, 78], insufficient consideration of energy consumption [38, 43, 52, 75], and a lack of clear energy model [19, 75].

5.3.2 Benchmark Datasets

In this work, two datasets are used to distinguish whether brain electrical activity either matches the preictal state or the interictal state using the AES [32] dataset and the CHB-MIT dataset [80].

The AES dataset includes iEEG recordings from five dogs and two human patients using an ambulatory monitoring system. The dataset consists of 48 seizures and 627.7 hours of interictal segments with naturally occurring epilepsy. These are long-duration recordings, spanning from multiple months and up to a year. More specifically, recordings from five dogs were sampled at 400 Hz, while those for the two patients were sampled at 5000 Hz. The recordings from four of the dogs are collected through 16 subdural electrodes and the remaining dog through 15 electrodes. One patient is recorded through 15 subdural electrodes, while the other is recorded through 24 electrodes. Also, each 10-minute annotated segment of the preictal and interictal state is extracted by medical experts. In this dataset, interictal recordings are enforced to be at least one week before or after any seizure. In contrast, preictal recordings cover one hour before seizure onset with a five-minute seizure prediction horizon.

The CHB-MIT dataset contains continuous sEEG recordings from 23 pediatric patients grouped into 24 cases using the international 10-20 system. Among them, recordings of chb01 and chb21 are the same female patient with a 1.5-year interval. The dataset consists of 198 seizures and total 940 hours of sEEG recordings captured after withdrawal of anti-seizure medication for up to several days. 23 sEEG signals have been recorded for most patients, but a few subjects have 24 or 26. All signals are sampled at 256 Hz. Several strategies are adopted in this work to meet the requirement of the seizure prediction problem. 21 common signal channels are selected to match patients with various sEEG signal channels, including FP1-F7, F7-T7, T7-P7, P7-O1, FP1-F3, F3-C3, C3-P3, P3-O1, FP2-F4, F4-C4, C4-P4, P4-O2, FP2-F8, F8-T8, T8-P8, P8-O2, FZ-CZ, CZ-PZ, T7-FT9, FT9-FT10 and FT10-T8. Interictal recordings must be at least 3 hours before and after seizure onset, while preictal recordings comprise 30 minutes before a seizure with a five-minute seizure prediction horizon.

The main difference between those two datasets is the type of EEG recordings, iEEG and sEEG respectively, which results in different spatial resolution and signal-to-noise ratio [81]. Compared with sEEG capturing signals through non-invasive electrodes affixed to the scalp, iEEG recording can obtain better spatial resolution and signal-to-noise ratio either via subdural grid and strip electrodes implanted directly on the cortical surface or through depth electrodes embedded into the brain parenchyma [82]. It is notable that g.Pangolin [83], reported the world's first ultra-high density non-invasive electrode grid to mitigate spatial resolution reduction with 1024 channels for sEEG recording. The non-invasive characteristics of sEEG still make it susceptible to noise and result in a lower signal-to-noise ratio.

5.3.3 Evaluation Metrics

As previously described, Figure 5.1 shows different brainwave patterns for epileptic patients. The seizure prediction problem can be considered as a binary classification between preictal and interictal states. Preictal durations of 60 and 30 minutes have been reported in [19,34] for the AES and the CHB-MIT datasets, respectively. For a reasonable comparison with recent state-of-the-art works, the same preictal duration is used for each dataset. After successful recognition of the preictal state, a 5-minute SPH allows medical intervention or mitigation measures before seizure onset. Moreover, the model's rigorous statistical evaluation methodology on two datasets was adopted. It is characterized by sensitivity, FPR, and the AUC through a five-fold cross-validation.

This work is also concerned with the energy required to perform the analysis which is highly dependent on memory access to data and model parameters. This is highly significant for wearable systems.

Statistical Evaluation

Sensitivity, FPR, and AUC are computed to evaluate our approach and to compare our results with state-of-the-art works. Sensitivity is the percentage of correctly classified 20-second seizure clips among the total 20-second. FPR is the false positive rate per hour [34]. AUC validates how good a given classifier is.

Energy Evaluation

This work aims to predict seizures with wearable systems with a small memory footprint. Thus, our energy model assumes on-chip memory access using a simplified parameterized energy model based on [71], which refers to a 28nm FDSOI fabrication technology. The weights and activations are prefetched into the static random-access memory (SRAM) buffer and directly fed into the multiply-accumulate (MAC) array for processing. As parallelism should only influence throughput without affecting the energy consumption for the same MAC area, weight-level and activation-level parallelism can be omitted [84]. To this end, the total on-chip energy consumption for inference consists of the following three parts: the computing energy E_C, weight access energy E_W, and activation access energy E_A.

$$\begin{aligned} E_C &= E_{MAC} \times (N_c + 3 \times A_s) \\ E_W &= E_M \times N_s \\ E_A &= 2 \times E_M \times A_s\,, \end{aligned} \tag{5.1}$$

where E_{MAC} is the energy consumed by MAC operations, E_M accounts for the energy consumed by a read/write from/to the SRAM, N_c is the number of MAC operations for partial product accumulation, N_s is expressed as the number of weights and biases within the model, and A_s is the number of activations. The energy of MAC operations is modeled as $E_{MAC} = 3.7pJ \times (b/16)^{1.25}$ [84], where b is the bit width used for quantized values. And E_M is modeled as $E_M = 2 \times E_{MAC}$. In addition, bias additions, BN, and activation operations are also assumed to cost E_{MAC}. Thus, the energy model adds $3 \times A_s$ to calculate E_C for the baseline model.

5.4 Methodology

5.4.1 Feature Extraction

The feature extraction includes data segmentation, resampling, and an average-pooling STFT. Each preictal or interictal EEG recording is cut into non-overlapping 20-second segments to augment the datasets. Next, each 20-second segment of the AES dataset is resampled at 400 Hz for convenient processing while keeping the 256 Hz sampling rate unchanged for the CHB-MIT dataset. To satisfy the Nyquist–Shannon sampling theorem, the maximum frequencies to avoid frequency aliasing for the AES dataset and the CHB-MIT dataset are 200 Hz and 128 Hz respectively. Due to the non-stationary feature of EEG signals, many complex feature extraction methods have been used to represent epileptic EEG signals, such as wavelet transform [33, 37, 85], STFT [34, 73], fractional Fourier transform [86]. Wavelet transform decomposes a signal into frequency components with varied time-frequency resolution, while STFT with fixed resolution means lower computation cost for battery-powered wearables. This work uses STFT to decompose time-series EEG signals into time-varying frequency items with rectangular windows. The ith segment of the STFT matrix $X(\omega) = [X_0(\omega), X_1(\omega), ..., X_{n-1}(\omega)]$ consists of the DFT of the windowed data centered at time point $(2i+1)\frac{L}{2}$ [66]:

$$\begin{aligned} X_i(\omega) &= \sum_{n=-\infty}^{\infty} x(n)w[n-(2i+1)\frac{L}{2}]e^{-j\omega n} \\ &= \mathrm{DTFT}_{\omega}(x \cdot \textsc{Shift}_{(2i+1)\frac{L}{2}}(w)) \,, \end{aligned} \tag{5.2}$$

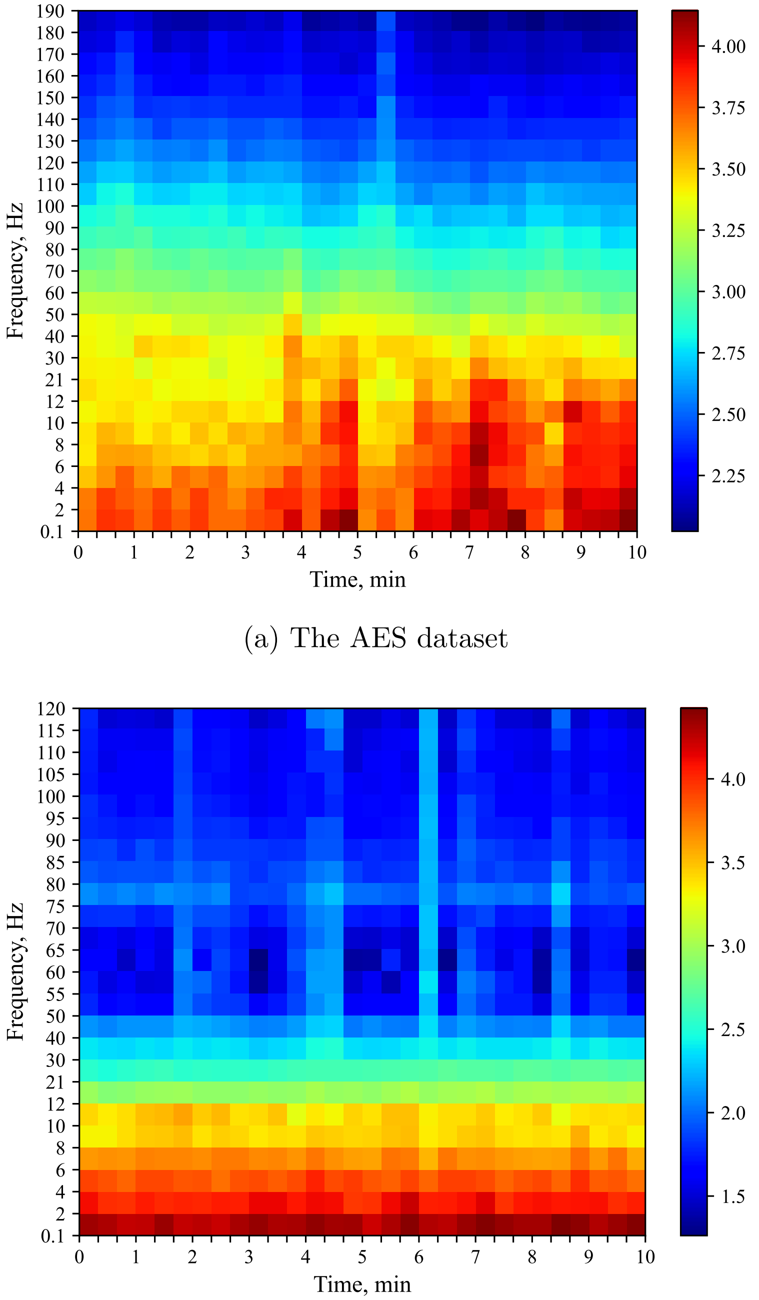
where $w[n-(2i+1)\frac{L}{2}]$ is the length L of the non-overlapping rectangular window function, which satisfies $\sum_i w[n-(2i+1)\frac{L}{2}] = 1$ for all $n \in \mathbb{Z}$. So the accumulation of all segments over time points corresponds to the DTFT of the whole signal $X(\omega)$:

$$\begin{aligned} \sum_{i=-\infty}^{\infty} X_i(\omega) &= \sum_{i=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} x(n)w[n-(2i+1)\frac{L}{2}]e^{-j\omega n} \\ &= \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \underbrace{\sum_{i=-\infty}^{\infty} w[n-(2i+1)\frac{L}{2}]}_{1 \text{ for all } n\in\mathbb{Z}} \\ &= \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \\ &= \mathrm{DTFT}_{\omega}(x) = X(\omega) \,. \end{aligned} \tag{5.3}$$

Furthermore, to improve the computational efficiency and to reduce the energy consumption of neural networks, the spectrum magnitude of average-pooling STFT can be calculated by the following function:

$$|X_i(k)| \;=\; \sum_{\omega\in\Omega_k} |X_i(\omega)| \,, \tag{5.4}$$

where Ω_k is the kth subset of the frequency bands. Since the brainwave frequencies are mainly concentrated in delta, theta, alpha, beta, and gamma frequency bands, Ω is divided into several corresponding frequency bands. More specifically, delta (0.1-2, 2-4 Hz), theta (4-6, 6-8 Hz), alpha (8-10, 10-12 Hz), beta (12-21, 21-30 Hz), low-gamma (30-40, 40-50, 50-60, 60-70, 70-80, 80-90, 90-100 Hz) and high-gamma (100-110, 110-120, 120-130, 130-140, 140-150, 150-160, 160-170, 170-180, 180-190 Hz) for the AES dataset [65,87]. To avoid frequency aliasing, the frequency ranges are acquired from 0.1 Hz to 190 Hz and from 0.1 Hz to 120 Hz through a fifth-order Butterworth band-pass filter for the AES dataset and the CHB-MIT dataset. To guarantee a minimum frequency resolution over 0.1 Hz, a 20-second rectangular window is applied with a frequency main lobe width of $\frac{2}{L}$ and a side lobe magnitude of -13dB. The raw 10-minute EEG signals are converted into corresponding frequency bands as shown in Figure 5.2.



(a) The AES dataset

(b) The CHB-MIT dataset

Figure 5.2 The spectrum magnitude of average-pooling STFT for 10-minute single-channel EEG preictal segments according to the two datasets

5.4.2 Neural Network Architecture

In this work, inspired by VGGNet [68] and one-dimensional CNN (1D CNN) [88], a tiny one-dimensional stacked CNN (1DSCNN) is manually designed and optimized to effectively predict epileptic seizures with low energy consumption while maintaining high performance. The VGGNet backbone is refined to focus on essential spatial features, while the 1D CNN contributes lightweight, energy-efficient spectrum processing over the temporal dimension. The employment of stacked convolutional layers, first introduced in the VGGNet architecture, offers two notable benefits: enhancing the depth of the neural network and decreasing the number of parameters while preserving the same receptive field to learn increasingly complex representations. Compared to two-dimensional CNNs or three-dimensional CNNs, the one-dimensional counterpart enables the extraction not only of pixel-level features within the image but also of finer details related to low-level characteristics such as edge shapes across multiple channels. Moreover, it allows the detection of trends in the spectrum distribution over the temporal dimension with low computational complexity. In addition, extensive hyperparameter tuning and layer adjustments are applied to ensure optimal performance. To this end, the proposed 1DSCNN leverages the advantages of VGGNet and 1D CNN, which offers an innovative and practical contribution for resource-limited wearable devices by balancing performance and energy efficiency.

The overall architecture of the proposed 1DSCNN is illustrated layer by layer in Figure 5.3. Initially, a 16-channel (AES) or 21-channel (CHB-MIT) spectrum of EEG signals is processed through a single 1D CNN block to extract cross-channel information simultaneously. This block is then followed by a 1D CNN block to form a complete 1DSCNN block. Subsequently, two additional 1DSCNN blocks are employed to enhance the model's generalization capability. ReLU activation functions and BN layers are utilized between each 1DSCNN block. Ultimately, a flatten layer and a dense layers are incorporated into a Softmax layer for classification.

5.4.3 Fixed-Precision Quantization

The computational flow with fixed-precision quantization for one 1DSCNN block is depicted in Figure 5.4. This quantization-aware training procedure, also well known as the simulated quantization, preserves the full-precision and fixed-precision values throughout the training process, which means that quantization effects are simulated through non-differentiable quantizers and quantized values are calculated in floating-point arithmetic [89]. After training, only the quantized weights, the quantization parameters such as scaling factors and zero points, and the parameters related to BN are retained for the inference procedure.

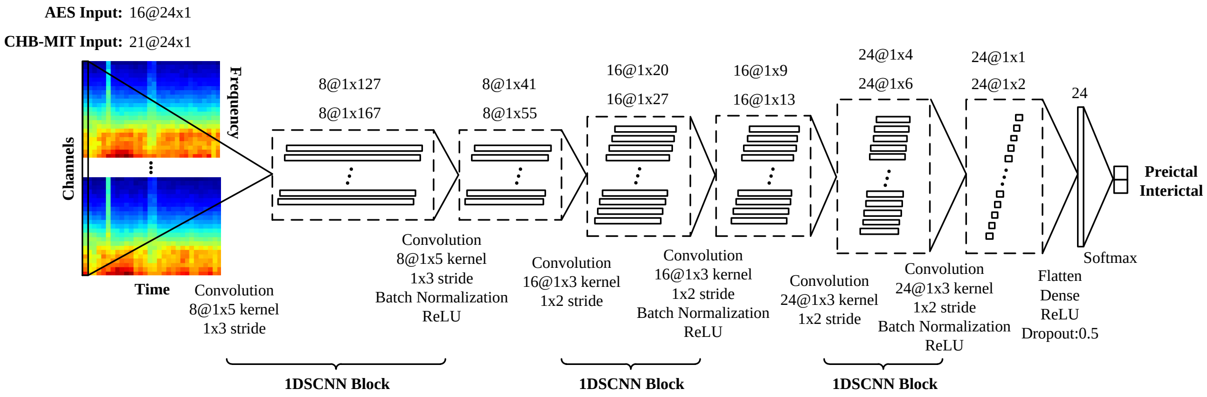


Figure 5.3 Architecture of the proposed convolutional neural network

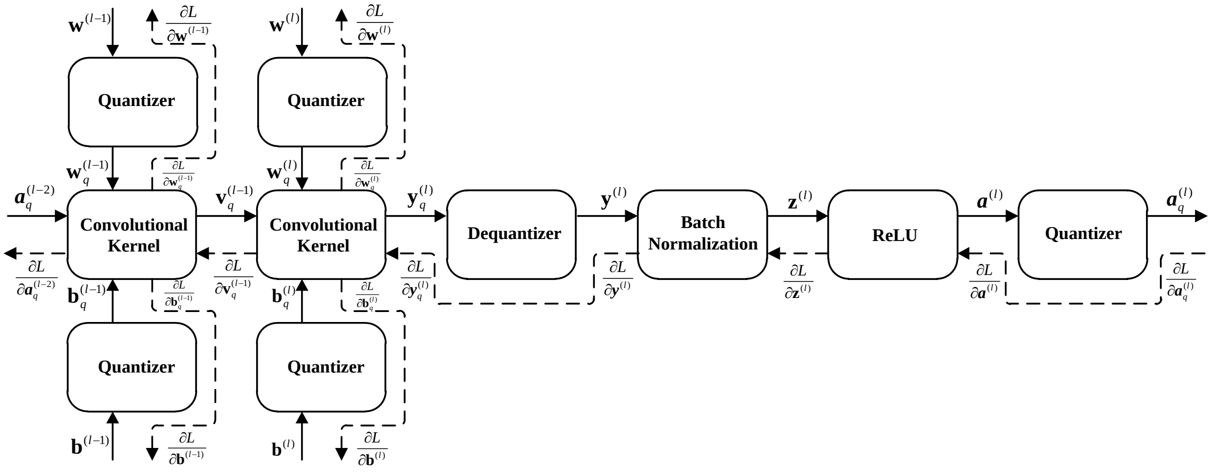

$\mathbf{w}^{(l-1)}$
$\frac{\partial L}{\partial \mathbf{w}^{(l-1)}}$
$\mathbf{w}^{(l)}$
$\frac{\partial L}{\partial \mathbf{w}^{(l)}}$
Quantizer
Quantizer
$\mathbf{w}_q^{(l-1)}$
$\frac{\partial L}{\partial \mathbf{w}_q^{(l-1)}}$
$\mathbf{w}_q^{(l)}$
$\frac{\partial L}{\partial \mathbf{w}_q^{(l)}}$
$\boldsymbol{a}_q^{(l-2)}$
Convolutional Kernel
$\mathbf{v}_q^{(l-1)}$
Convolutional Kernel
$\mathbf{y}_q^{(l)}$
Dequantizer
$\mathbf{y}^{(l)}$
Batch Normalization
$\mathbf{z}^{(l)}$
ReLU
$\boldsymbol{a}^{(l)}$
Quantizer
$\boldsymbol{a}_q^{(l)}$
$\frac{\partial L}{\partial \boldsymbol{a}_q^{(l-2)}}$
$\mathbf{b}_q^{(l-1)}$
$\frac{\partial L}{\partial \mathbf{b}_q^{(l-1)}}$
$\frac{\partial L}{\partial \mathbf{v}_q^{(l-1)}}$
$\mathbf{b}_q^{(l)}$
$\frac{\partial L}{\partial \mathbf{b}_q^{(l)}}$
$\frac{\partial L}{\partial \mathbf{y}_q^{(l)}}$
$\frac{\partial L}{\partial \mathbf{y}^{(l)}}$
$\frac{\partial L}{\partial \mathbf{z}^{(l)}}$
$\frac{\partial L}{\partial \boldsymbol{a}^{(l)}}$
$\frac{\partial L}{\partial \boldsymbol{a}_q^{(l)}}$
Quantizer
Quantizer
$\mathbf{b}^{(l-1)}$
$\frac{\partial L}{\partial \mathbf{b}^{(l-1)}}$
$\mathbf{b}^{(l)}$
$\frac{\partial L}{\partial \mathbf{b}^{(l)}}$

Figure 5.4 Computation flow with fixed-precision quantization for one 1DSCNN block. The solid lines indicate the forward pass, while the dashed lines represent the backward pass

In the forward pass, starting from a floating-point vector $\mathbf{x} = \{x_1, \cdots, x_N\}$ with range (x_{min}, x_{max}), the dequantized value $\widehat{\mathbf{x}}$ with b-bit-width precision can be represented as [54]

$$\begin{aligned} \widehat{\mathbf{x}} &= q(\mathbf{x}; s, z, p, q) \\ &= \underbrace{s\Bigg[\underbrace{\text{clamp}\left(\left\lfloor \frac{\mathbf{x}}{s} \right\rceil + z; p, q\right)}_{\text{Quantizer}} - z\Bigg]}_{\text{Dequantizer}}, \end{aligned} \tag{5.5}$$

where the clipping thresholds are set as $[p, q] = [0, 2^b - 1]$ for unsigned integer mapping and $[p, q] = [-2^b, 2^b - 1]$ for signed integer mapping, $\lfloor \cdot \rceil$ indicates the round-to-nearest function, the clamp$(\cdot)$ function restricts all values to be within the range p to q, s is the scaling factor and z is the bias value relative to the zero point.

This work uses a uniform-signed symmetric quantizer for the weights and biases, while a uniform-signed asymmetric quantizer is employed for the activations. The flow of quantizers and dequantizers is illustrated in Figure 5.4. To improve computational efficiency for inference, the scaling factor s is a power of two, illustrating that bit-shifting operations can replace multipliers. The scaling factor for the weights and biases is demonstrated as

$$s_{sym} = 2^{-\left\lfloor \log_2 \frac{p}{\max(|x_{max}|, |x_{min}|)} \right\rceil}. \tag{5.6}$$

To cover the entire vector range, Min-max scaling is applied to the activations as

$$s_{asym} = 2^{-\left\lfloor \log_2 \frac{p-q}{x_{max} - x_{min}} \right\rceil}, \tag{5.7}$$

which results in no clipping errors and acceptable rounding errors when there are no outliers. Meanwhile, the scaling factor is determined per layer for the activations and biases, while it is set per channel for the weights [90]. Notably, the clipping threshold for the weights and biases is set as $[p, q] = [-2^b + 1, 2^b - 1]$, because the quantized weights and biases can not only be maintained in symmetry ($z = 0$) but also mitigate the risk of overflow after a multiply-add operation for inference [89]. In addition, z of an asymmetric quantizer for the activations is shown as

$$z_{asym} = \text{clamp}\left(-\left\lfloor \frac{x_{max} + x_{min}}{2s_{asym}} \right\rceil; -2^b + 1, 2^b - 1\right). \tag{5.8}$$

In the backward pass, a STE is used for gradient approximation [54, 91], then the gradient can be calculated through the chain rule as

$$\begin{aligned}\frac{\partial \widehat{\mathbf{x}}_i}{\partial \mathbf{x}_i} &= \frac{\partial q(\mathbf{x}_i; s, z, p, q)}{\partial \mathbf{x}_i} \\ &= s \cdot \frac{\partial}{\partial \mathbf{x}_i} \operatorname{clamp}\left(\left\lfloor \frac{\mathbf{x}_i}{s} \right\rceil; p, q\right) + 0 \\ &= \begin{cases} 1, & \text{if } p \leq \left\lfloor \frac{\mathbf{x}_i}{s} \right\rceil \leq q, \\ 0, & \text{otherwise,} \end{cases}\end{aligned} \tag{5.9}$$

the non-differentiable quantizer function is skipped as described in Figure 5.4.

5.4.4 Mixed-Precision Quantization

In mixed-precision quantization, the purpose is to apply low-bit-width fixed-point quantization to insensitive or less sensitive layers, while keeping sensitive and efficient layers in high-bit-width fixed-point resolution. Thus, each layer is quantized through various bit-width for searching the combination offering the best performance. An exhaustive search through all possible bit-width combinations is very expensive or not feasible as the number of possible solutions grows exponentially with the problem size. Wang et al. [92] designed an automatic quantization policy by getting the feedback of reinforcement learning agents from the hardware simulator. Wu et al. [93] explored the mixed-precision search problem with the differentiable neural architecture search method to effectively address the search space. Although these exhaustive search methods can find the optimal solution, there is a high computational cost associated with them due to the time complexity of $\mathcal{O}(L^b)$, where L is the number of layers and b is the number of bit widths being considered.

To speed up the search and to reduce computational complexity, a near-optimal low-complexity mixed-precision search procedure is proposed as shown in Algorithm 1, with a time complexity of $\mathcal{O}(L \times b)$. The ultimate goal is to find the mixed-precision model m^* by layerwise bit-width search. Given the mixed-precision model's sensitivity to activations, a consistent 8-bit precision for all layer activations is implemented without compromising performance [94]. The layer-wise bit-width vector is defined as $b = [b_1, b_2, ..., b_L]$, where each element b_i specifies the number of bits used to represent model weights and biases in layer i. To further simplify the search, we constrain $b_i \in \{32, 16, 8, 4, 2, 1\}$. The $\textsc{train}(s)$ procedure corresponds to training the model from scratch, with the weight and bias representation of each layer determined by vector s, that is layer i will be assigned a weight and bias representation with s_i bits. Functions AUC(m) and SIZE(m) calculate the AUC and size of model m, respectively.

Following standard practice, the first and last layers in the model use higher precision weights

Algorithm 1: Mixed-precision search procedure

```
   output: Mixed-precision model $m^*$
1  begin
2      $b_1 \leftarrow 32$
3      $b_L \leftarrow 32$
4      for $i \leftarrow 2$ to $L - 1$ do
5          $s \leftarrow [b_1, 16, 16, \ldots, 16, b_L]$
6          $M \leftarrow \{\}$
7          for $n \in \{8, 4, 2, 1\}$ do
8              $s_i \leftarrow n$
9              $m \leftarrow \textsc{train}(s)$
10             add $(m, n)$ to set $M$
11         end
12         $M \leftarrow \arg\max_{(m,n)\in M} \text{AUC}(m)$
13         $(m, n) \leftarrow \arg\min_{(m,n)\in M} \text{SIZE}(m)$
14         $b_i \leftarrow n$
15     end
16     $m^* \leftarrow \textsc{train}(b)$
17 end
```

and biases. A layer-wise bit-width search is then performed by looping from layer 2 to layer $L - 1$ (line 4). Firstly, the default value of the layer-wise bit-width vector s is set at 16 for the $L - 2$ remaining layers (excluding the first layer and the last layer) (line 5). Then, while looping s_i over $\{8, 4, 2, 1\}$ for a specific layer i (line 7, 8), the model m is trained according to configuration s (line 9) and the result is saved with corresponding bit-width to set M (line 10). Next, the set of models that maximize the average AUC (there could be multiple such models) is again saved in M (line 12). From this new set M, we retain the model with the smallest average model size (line 13). Note that only one model will have a minimum size while maximizing the AUC, since all models in M have different sizes. Finally, the optimal bit-width is saved in b_i for ith layer (line 14). The search procedure (lines 4-15) is repeated until the layer-wise bit-width vector b has been determined. Then, the mixed-precision model m^* is trained according to the determined b.

5.5 Experimental Results

5.5.1 Training Settings

For each subject, the annotated two datasets are randomly split into 80% training set and 20% test set through a five-fold stratified cross-validation [95]. Then, 20% of the training set is allocated to the validation set to prevent overfitting during the training phase with an early stopping criterion of accuracy [95]. The patient-specific model is trained for each subject to achieve the best performance in the validation set, and validated in the test set later. As the distribution of the preictal and the interictal state in epileptic seizure prediction is extremely skewed [77], inversely proportional class weights of the number of the preictal and the interictal state are assigned to the loss function during the training procedure. The general training strategy is configured with a dropout rate of 0.5 and a batch size of 16. For the baseline model, the Adam optimizer is adopted in training 100 epochs, with β_1 of 0.9, β_2 of 0.999, a learning rate decayed every 50 epochs by half starting at 0.001 for the AES dataset, and decayed every 25 epochs by half starting at 0.01 for the CHB-MIT dataset. For the quantization model, the Adam optimizer is also utilized for training with the same β_1 and β_2 values, except 200 epochs and a learning rate decayed every 50 epochs by half starting at 0.001 for both two datasets. In particular, no quantization is applied to the input and output layers.

5.5.2 Results and Discussion

Performance Evaluation and Analysis

For the 32-bit floating-point baseline, our proposed patient-specific 1DSCNN models are applied to all subjects of the two datasets and evaluated using statistical evaluation metrics such as sensitivity, FPR, and AUC through a five-fold cross-validation. The average results are reported in Table 5.1 and Table 5.2, the corresponding AUC box plots for each subject are shown in Figure 5.5 and Figure 5.6. For the AES dataset, as reported in Table 5.1, an average sensitivity of 94.44%, an average FPR of 0.011/h, and an average AUC of 0.979 are achieved, as previously reported in [87]. For the CHB-MIT dataset, as reported in Table 5.2, an average sensitivity of 96.14%, an average FPR of 0.018/h, and an average AUC of 0.996 are achieved, which validates the generalization ability of our neural network architecture. The AUC box plots per subject for the two datasets are presented in Figure 5.5 and Figure 5.6, where the yellow line in the box represents the median AUC value, the edges of the box indicate the first and third quartile edge values of the AUC and the bar that extends off the box goes from minimum to maximum observed AUC values. It is noteworthy that dog_1 and chb14 are the most challenging subjects for seizure prediction, as evidenced by the persistently lowest AUC values, even after hyperparameter tuning.

Then, to further reduce the model size for low-power wearables, energy-efficient quantization is applied to our single-precision floating-point baseline models as presented in Table 5.3 and Table 5.4. A trend of performance degradation is observed as the levels of model quantization

Table 5.1 Per-subject statistical evaluation results for the AES dataset: sensitivity, FPR, and AUC

Subject	Sensitivity(%)	FPR(/h)	AUC
dog_1	91.11	0.013	0.926
dog_2	97.70	0.001	0.998
dog_3	95.42	0.003	0.978
dog_4	92.27	0.003	0.974
dog_5	96.78	0.001	0.999
patient_1	97.22	0.008	0.998
patient_2	90.56	0.045	0.979
Average	**94.44**	**0.011**	**0.979**

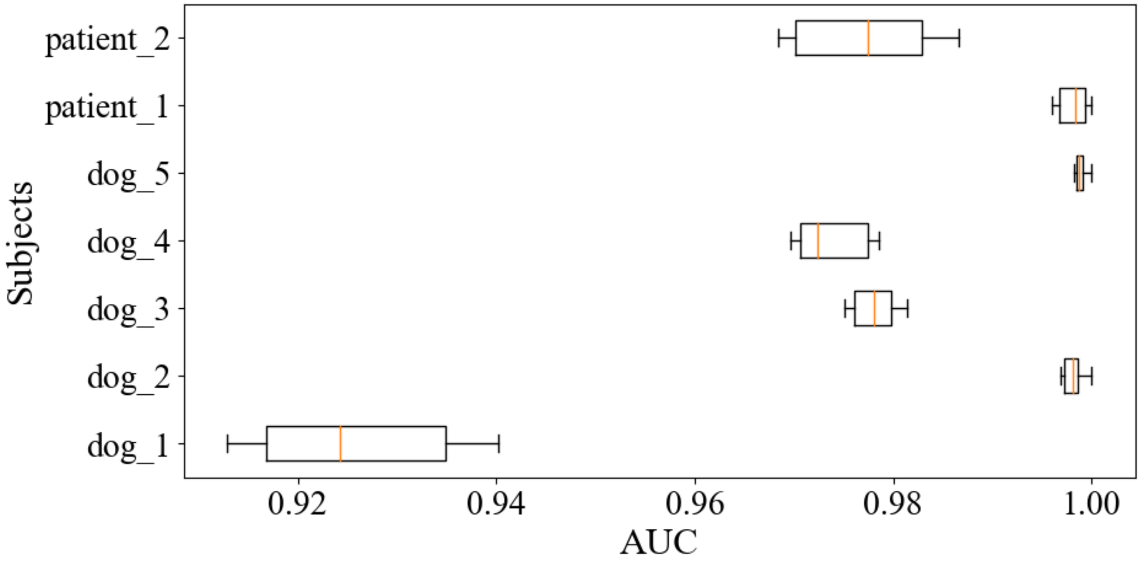


Figure 5.5 Per-subject AUC box plot for the AES dataset: median, quartile, and extreme deviations of observed AUC scores obtained from a five-fold cross-validation

Table 5.2 Per-subject statistical evaluation results for the CHB-MIT dataset: sensitivity, FPR and AUC

Subject	Sensitivity(%)	FPR(/h)	AUC
chb01	99.21	0.001	1.000
chb02	99.63	0.000	1.000
chb03	96.85	0.001	0.998
chb04	95.19	0.000	1.000
chb05	87.56	0.011	0.982
chb06	97.41	0.001	1.000
chb07	96.30	0.000	0.998
chb08	98.89	0.019	0.999
chb09	98.33	0.000	1.000
chb10	93.89	0.004	0.996
chb11	93.33	0.001	0.995
chb12	98.10	0.056	0.997
chb13	98.22	0.002	0.999
chb14	89.11	0.084	0.971
chb15	97.53	0.125	0.994
chb16	93.70	0.022	0.993
chb17	98.89	0.000	1.000
chb18	97.04	0.001	0.998
chb19	94.81	0.000	1.000
chb20	98.67	0.002	0.999
chb21	95.00	0.001	0.998
chb22	92.78	0.002	0.997
chb23	97.56	0.006	0.999
chb24	99.35	0.090	0.998
Average	**96.14**	**0.018**	**0.996**

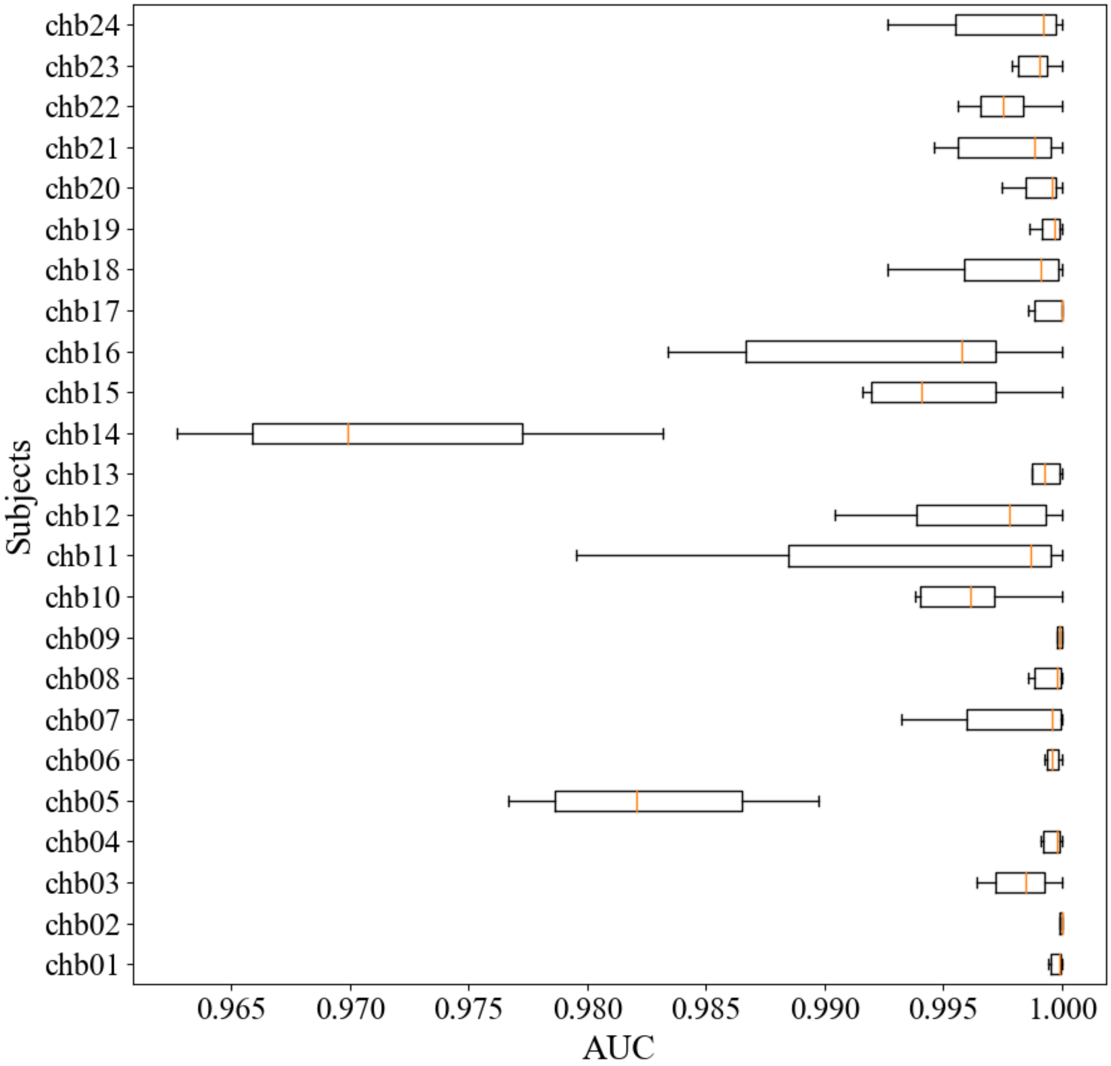


Figure 5.6 Per-subject AUC box plot for the CHB-MIT dataset: median and deviation of AUC scores by a five-fold cross-validation

Table 5.3 Per-subject AUC scores and model sizes for the AES dataset: single-precision floating point, fixed-precision quantization from 16-bit to 1-bit, and mixed-precision quantization

Subject	**FP**	**Quantization**					
		16-bit	**8-bit**	**4-bit**	**2-bit**	**1-bit**	**MP**
dog_1	0.926	0.923	0.924	0.911	0.875	0.763	0.926
dog_2	0.998	0.998	0.998	0.997	0.984	0.938	0.997
dog_3	0.978	0.978	0.977	0.973	0.960	0.858	0.975
dog_4	0.974	0.974	0.970	0.970	0.936	0.868	0.971
dog_5	0.999	0.999	0.999	0.999	0.995	0.956	0.999
patient_1	0.998	0.999	0.998	0.997	0.991	0.973	0.998
patient_2	0.979	0.978	0.975	0.974	0.950	0.875	0.970
Average	**0.979**	**0.978**	**0.977**	**0.974**	**0.956**	**0.890**	**0.977**
Model size (kB)	**21.32**	**10.86**	**5.62**	**3.01**	**1.70**	**1.05**	**3.02**

evolve from 16-bit to 1-bit, but is also accompanied by a reduction in memory footprint and computational complexity. Hence, there is a trade-off between energy consumption and performance for each subject. Furthermore, the extent of performance degradation varies across different subjects in different datasets, which means that different subjects exhibit varying sensitivities to distinct levels of quantization. Notably, compared to the 32-bit floating-point baseline, 8-bit quantization reduces the model size by 3.79 times with only a 0.20% average AUC score loss for the AES dataset, while 4-bit quantization reduces the model size by 5.95 times with the same 0.20% average AUC score loss for the CHB-MIT dataset. It appears that the AES dataset is more sensitive to quantization than the CHB-MIT dataset. However, it deserves attention that applying 16-bit quantization reduces the model size by half for both datasets, with an average AUC score decrease of 0.10% for the AES dataset and 0.20% for the CHB-MIT dataset. The primary reason for this phenomenon is that quantization errors, including rounding and clipping errors (weights, biases, and scaling factors), initially lead to performance degradation. Then, as the model is further quantized, these errors may paradoxically contribute to performance improvement through a regularization effect [96]. Evidently, the CHB-MIT dataset's signal-to-interference ratio surpasses that of the AES dataset from Figure 5.2. It demonstrates that, for the CHB-MIT dataset, quantization errors initially dominate and lead to performance degradation. However, as the model is further quantized, these errors begin to contribute to performance improvement.

Table 5.4 Per-subject AUC scores and model sizes for the CHB-MIT dataset: single-precision floating point, fixed-precision quantization from 16-bit to 1-bit, and mixed-precision quantization

Subject	**FP**	**Quantization**					
		16-bit	**8-bit**	**4-bit**	**2-bit**	**1-bit**	**MP**
chb01	1.000	0.999	1.000	1.000	0.999	0.988	0.999
chb02	1.000	1.000	1.000	1.000	1.000	0.998	1.000
chb03	0.998	0.998	0.998	0.998	0.995	0.963	0.997
chb04	1.000	1.000	1.000	1.000	0.997	0.983	1.000
chb05	0.982	0.967	0.964	0.969	0.929	0.903	0.966
chb06	1.000	1.000	0.999	0.999	0.998	0.988	0.999
chb07	0.998	0.999	0.999	0.998	0.996	0.982	0.997
chb08	0.999	0.999	0.998	0.997	0.997	0.978	0.999
chb09	1.000	1.000	0.999	0.999	1.000	0.995	0.999
chb10	0.996	0.995	0.995	0.995	0.985	0.943	0.988
chb11	0.995	0.993	0.988	0.990	0.980	0.947	0.990
chb12	0.997	0.996	0.997	0.995	0.997	0.980	0.997
chb13	0.999	0.999	0.999	0.999	0.998	0.988	0.997
chb14	0.971	0.955	0.951	0.960	0.932	0.822	0.971
chb15	0.994	0.991	0.991	0.989	0.989	0.952	0.990
chb16	0.993	0.988	0.987	0.983	0.972	0.914	0.985
chb17	1.000	1.000	1.000	1.000	1.000	0.999	1.000
chb18	0.998	0.998	0.998	0.996	0.996	0.974	0.996
chb19	1.000	0.998	0.997	0.997	0.995	0.984	0.999
chb20	0.999	1.000	0.999	1.000	1.000	0.995	0.999
chb21	0.998	0.996	0.997	0.996	0.995	0.980	0.996
chb22	0.997	0.994	0.996	0.996	0.997	0.982	0.996
chb23	0.999	0.997	0.998	0.996	0.996	0.977	0.997
chb24	0.998	0.998	0.999	0.996	0.993	0.984	0.998
Average	**0.996**	**0.994**	**0.994**	**0.994**	**0.989**	**0.967**	**0.993**
Model size (kB)	**23.62**	**12.39**	**6.78**	**3.97**	**2.56**	**1.86**	**2.75**

In contrast, for the AES dataset, quantization errors improve performance from the outset. This phenomenon becomes more apparent for mixed-precision quantization, particularly for subjects that are challenging to predict, such as dog_1 and chb14. A very significant result is that mixed-precision quantization contributes additional regularization, potentially enhancing model performance to a certain extent, and may even yield AUC results comparable to those achieved with a 32-bit floating-point model [97] while reducing the model size by 7.06 and 8.59 times for the AES and CHB-MIT data sets respectively.

Comparative Analysis and Discussion

For a robust comparative analysis, it is crucial to consider factors such as dataset consistency, feature types, sampling segmentation, preictal duration, and training methodologies, which can substantially influence model performance. Therefore, Table 5.5 presents a comparative analysis of this work with other leading epileptic seizure prediction methods. All methods in Table 5.5 employ CNNs either fully or partially for classification, except [43, 52], which use LSTM network and B2-ViT Net. In addition, all methods are patient-specific, except for [79], which incorporates cross-patient generalization. Regarding the AES dataset, the majority of methods [19, 38, 52, 61, 79] adopt a 60-minute preictal duration, whereas for the CHB-MIT dataset, a 30-minute duration is predominantly utilized [19,34,51,52,75,77–79]. In particular, models previously mentioned in [19,38,61] are rebuilt to determine their respective sizes, which subsequently facilitates the comparison of energy consumption per inference with this work. In terms of the AES dataset, the proposed baseline model (1DSCNN) outperforms prior approaches, achieving the best sensitivity, FPR, and AUC while maintaining the lowest model size and energy consumption. Furthermore, compared to the baseline model, 8-bit quantization reduces the model size by 3.79 times and energy consumption by half with merely a 0.20% average AUC score loss, while mixed-precision quantization reduces the model size by 7.06 times and energy consumption by 2.22 times with the same 0.20% average AUC score loss. Then, regarding the CHB-MIT dataset, it should be noticed that various methods adopt different numbers of patients. Consequently, the top 10 patients, distinguished by the highest AUC scores, are selected from the 24 patients to allow for a fairer comparative analysis of performance with [19,43]. The proposed baseline model (24 patients) achieves the best sensitivity, FPR, AUC with the lowest model size when compared to [34,51,52,75,77–79]. Compared to the baseline model, 4-bit quantization results in a 5.95-time reduction in model size and a 2.33-time reduction in energy consumption with only a 0.20% average AUC score loss, while mixed-precision quantization results in an 8.59-time reduction in model size and 2.33-time reduction in energy consumption with the 0.30% average AUC score loss. Moreover, the proposed baseline model (10 patients) achieves sensitivity, FPR, and AUC comparable

Table 5.5 Comparison to prior state-of-the-art methods

Method	**Dataset**	**# of subjects**	**Features**	**Classifier**	**Sensitivity (%)**	**FPR (/h)**	**AUC**	**Model size**	**Energy (μJ)**	**Preictal duration**
Korshunova et al. 2017 [38]	AES	5 dogs, 2 patients	Spectral power, signal standard deviation	CNN	-	-	0.810	0.56 MB	1.24	60 min
Liu et al. 2019 [61]	AES	5 dogs, 2 patients	PCA, mean log spectral power	Multi-view CNN	-	-	0.837	1.79 MB	9.43	60 min
Daoud et al. 2019 [43]	CHB-MIT	8 patients	DCAE	Bi-LSTM	99.72	0.004	-	71.66 kB	-	60 min
Zhang et al. 2019 [77]	CHB-MIT	23 patients	Combination of common spatial pattern	CNN	92.2	0.12	0.90	-	-	30 min
Tang et al. 2020 [78]	CHB-MIT	24 patients	Fractal spectrum, relative band energy, PLV modularity	Multi-view CGRN	94.5	0.118	0.901	-	-	30 min
Wu et al. 2023 [75]	CHB-MIT	11 patients	Raw data	CNN based ResNet	94.2	0.09	-	-	-	30 min
Lee et al. 2024 [51]	CHB-MIT	24 patients	Short-time Fourier transform	Pre-train+ ResNet-LSTM	81.54	0.073	-	-	-	30 min
Truong et al. 2018 [34]	AES	5 dogs, 2 patients	Short-time Fourier transform	CNN	75	0.21	-	0.76 MB	-	30 min
	CHB-MIT	13 patients			81.2	0.16	-	0.77 MB	-	30 min
Zhao et al. 2021 [19]	AES	5 dogs, 2 patients	Raw data	CNN	93.48	0.063	0.977	44.52 kB	153.82	60 min
	CHB-MIT	10 patients			99.81	0.005	1	45.22 kB	141.46	30 min
Liang et al. 2023 [79]	AES	5 dogs	Short-time Fourier transform	Semi-supervised domain adaptation	75.7	0.165	0.763	-	-	60 min
	CHB-MIT	13 patients			88.8	0.182	0.849	-	-	30 min
Shi et al. 2024 [52]	AES	5 dogs	Short-time Fourier transform	B2-ViT Net	85.2	0.013	0.816	19.07 MB	-	60 min
	CHB-MIT	13 patients			93.3	0.057	0.923	19.08 MB	-	30 min
This work	**AES**	**5 dogs, 2 patients**	**Average-pooling short-time Fourier transform**	**1DSCNN**	**94.44**	**0.011**	**0.979**	**21.32 kB**	**0.20**	**60 min**
				8-bit Quant.	**92.49**	**0.011**	**0.977**	**5.62 kB**	**0.10**	**60 min**
				MP Quant.	**91.75**	**0.013**	**0.977**	**3.02 kB**	**0.09**	**60 min**
	CHB-MIT	**10 patients**		**1DSCNN**	**97.92**	**0.003**	**1**	**23.62 kB**	**0.28**	**30 min**
		24 patients		**1DSCNN**	**96.14**	**0.018**	**0.996**	**23.62 kB**	**0.28**	**30 min**
				4-bit Quant.	**94.57**	**0.023**	**0.994**	**3.97 kB**	**0.12**	**30 min**
				MP Quant.	**94.22**	**0.022**	**0.993**	**2.75 kB**	**0.12**	**30 min**

Table 5.6 The paired t-test of model size efficiency between mixed-precision quantization and single-precision floating point, representative fixed-precision quantization on the AES and CHB-MIT datasets

Dataset	**Comparison**	**p-value**
AES	MP vs FP	< 0.001
	MP vs 8-bit	< 0.001
CHB-MIT	MP vs FP	< 0.001
	MP vs 4-bit	< 0.001

Table 5.7 The paired t-test of AUC between various precisions and two recent methods on the AES and CHB-MIT datasets, with values greater than 0.05 highlighted in bold

Dataset	**Method**	**FP**	**Quantization**					
			16-bit	**8-bit**	**4-bit**	**2-bit**	**1-bit**	**MP**
AES	Liang et al. [79]	0.002	0.002	0.002	0.001	< 0.001	0.007	0.002
	Shi et al. [52]	0.024	0.023	0.024	0.023	0.022	**0.119**	0.025
CHB-MIT	Liang et al. [79]	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	Shi et al. [52]	0.009	0.011	0.010	0.011	0.016	**0.074**	0.011

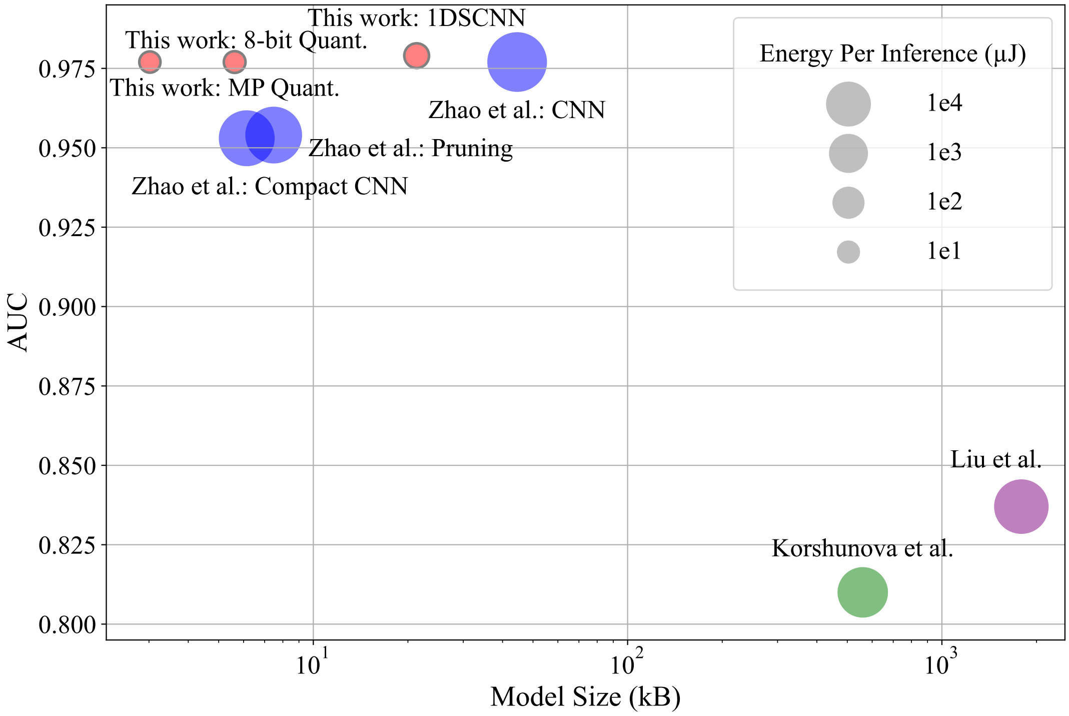


Figure 5.7 Comparison of energy consumption per inference for the AES dataset: the size of the gray balls depicted in the legend corresponds to the relative energy consumption

to [19, 43], while reducing model size by 3.03 and 1.91 times, respectively, leading to lower energy consumption.

Table 5.6 presents the paired t-test results at a significance level of 0.05 for the model size efficiency between mixed-precision quantization and single-precision floating point, representative fixed-precision quantization on the AES and CHB-MIT datasets, respectively. This table shows the significance of the ablation experiments, which include four comparisons: mixed-precision quantization vs. full precision and mixed-precision quantization vs. 8-bit quantization for the AES dataset, mixed-precision quantization vs. full precision and mixed-precision quantization vs. 4-bit quantization for the CHB-MIT dataset. Statistical analysis reveals that the proposed mixed-precision approaches significantly outperform both full-precision and representative fixed-precision quantization methods on the AES and CHB-MIT datasets. Table 5.7 demonstrates the paired t-test results at a significance level of 0.05 for the AUC values between the proposed models of various precisions and two recent methods on the AES and CHB-MIT datasets, respectively. This table shows that all our methods outperform the recent methods proposed in [79] and [52] on both datasets, with statistically significant differences ($p < 0.05$), except for the 1-bit quantization when compared to the method in [52].

Figure 5.7 shows a comparison of average AUC, model size, and energy consumption per inference between this work and [19, 38, 61] for the AES dataset. It should be noted that Zhao et al. [19] use raw data as input for CNNs, in contrast to other methods that employ preprocessing techniques. The weights and activations are preloaded into the on-chip SRAM buffer to calculate energy consumption. In addition, quantization is not applied to either the input or the output layers. To this end, this work, including both the baseline and quantization models, significantly outperforms other methods in terms of prediction performance and energy consumption, which demonstrates exceptional potential for low-power biomedical wearable devices.

5.6 Conclusion

In this work, a tiny neural network has been proposed for patient-specific epileptic seizure forecasting in wearable devices and validated on the AES dataset and the CHB-MIT dataset. The proposed baseline method is validated through a five-fold stratified cross-validation and outperforms prior state-of-the-art methods in terms of sensitivity, FPR, AUC, model size, and energy consumption for both datasets. The baseline method achieves an average sensitivity of 94.44%, an average FPR of 0.011/h, an average AUC of 0.979, a model size of 21.32 kB, and an estimated energy consumption of 0.20 μJ for the AES dataset. Meanwhile, it attains an average sensitivity of 96.14%, an average FPR of 0.018/h, an average AUC of 0.996, a model size of 23.62 kB, and an estimated energy consumption of 0.28 μJ for the CHB-MIT dataset. Subsequently, to more effectively cater to the requirements of low-power wearables, the baseline method is enhanced with hardware-friendly fixed-precision and mixed-precision quantization. This approach serves to further improve the energy efficiency of our baseline method by at least two times. Notably, our mixed-precision quantization can potentially improve the model performance of hard-to-predict subjects, such as dog_1 and chb14, due to additional regularization. To this end, our proposed tiny neural network was shown to be hardware-friendly, with great potential to facilitate the development of biomedical wearable devices.

CHAPTER 6 ARTICLE 3: $S^3$1DCNN: A COMPACT STACKED SPECTRAL-SPATIAL ATTENTION 1DCNN FOR SEIZURE PREDICTION WITH WEARABLES

Yang Zhang[1], Yvon Savaria[1], Mohamad Sawan[1,2], François Leduc-Primeau[1]
[1]Department of Electrical Engineering, Polytechnique Montreal, Canada
[2]School of Engineering, Westlake University, Hangzhou, Zhejiang, China

Published in: the 22st IEEE Interregional NEWCAS Conference (NEWCAS)
Publication date: September 17, 2024

The third objective is to develop a high-performance, interpretable CNN with small energy consumption proposed for precisely forecasting epileptic seizures, intended to be transparent and understandable to healthcare professionals. To achieve the third objective, a seizure prediction method using an attention-driven $S^3$1DCNN is introduced, showcasing its capability to interpret and analyze spatio-temporal, non-stationary iEEG recordings for accurate identification of epilepsy onset regions. The spectral-spatial attention can merge spectral and spatial information within its local receptive field to create enhanced feature representations by dynamically weighting the significance of various features. Furthermore, the proposed $S^3$1DCNN surpasses state-of-the-art methods in terms of AUC, model size, and energy consumption on the AES dataset, which demonstrates its great potential for low-power biomedical wearable devices. In contrast to the methods in Chapter 4 and Chapter 5, the attention-driven $S^3$1DCNN introduced in this chapter results in a slight increase in model size and energy consumption, yet it marginally outperforms them and offers interpretability to clarify the location of the seizure onset. Chapter 7 will give more substance and provide detailed information on the costs and benefits compared to the other chapters.

6.1 Abstract

Seizure prediction has become a crucial field of research that aims to improve the lives of patients with drug-resistant epilepsy by reducing their anxiety and allowing the implementation of precautionary measures. Recently, deep learning has shown remarkable advancements in epilepsy prediction. However, this progress comes with increased computational demands and memory usage, which makes it unsuitable for low-power wearable devices. This work proposes a compact $S^3$1DCNN leveraging the STFT. This model aims to enhance the interpretable ability to analyze non-stationary EEG signals, making it suitable for implementation

in wearable biomedical devices. The results demonstrate that the proposed method outperforms recent state-of-the-art methods, achieving an average sensitivity of 92.1%, an average FPR of 0.008/h, an average AUC of 0.980, and an estimated energy consumption of 0.21 μJ per inference on the AES dataset. It demonstrates our method's promising application potential in low-power and energy-efficient wearable devices.

6.2 Introduction

Epilepsy is a chronic neurological disorder characterized by recurrent, unprovoked seizures, which affects approximately 50 million individuals across all age groups [16]. Treatment for epilepsy typically relies on anti-epileptic medication. However, long-term medication therapy can induce a series of side effects, among which medication resistance ranks as one of the most challenging issues. Therefore, it is particularly crucial for individuals with epilepsy to anticipate the onset of seizures, enabling better management of medication to minimize side effects and optimize dosage. To this end, seizure prediction is vital in restoring their quality of life.

A conceptual framework for epilepsy seizure prediction in wearable biomedical devices is shown in Figure 6.1. iEEG recordings are obtained through sensors implanted within the brain, which then wirelessly transmit the data to an external wearable biomedical device for analysis. In this work, seizure prediction is viewed as a binary classification [19, 34, 38, 61, 65], distinguishing between preictal states (pre-seizure period) and interictal states (seizure-free period) as in Figure 6.1. The SPH is the interval between the preictal and ictal state (seizure onset), during which it is optimal for deploying medical interventions.

Although recent advances in deep learning techniques have led to substantial improvements in the performance of seizure prediction [19, 34, 38, 61, 65], it is essential to account for more practical details such as power consumption for implanted sensors and wearable devices, the trade-off between power consumption and performance. To achieve this, a compact and hardware-friendly CNN designed for epileptic seizure prediction is proposed. Our main contributions are summarized as follows:

1. An efficient data processing and compression method, specifically designed for devices with limited energy resources, such as implanted devices, is presented.

2. A compact $S^3$1DCNN is proposed to improve the interpretability of non-stationary EEG signals for wearables. It surpasses existing statistical evaluation methods and achieves the lowest energy consumption among comparable methods.

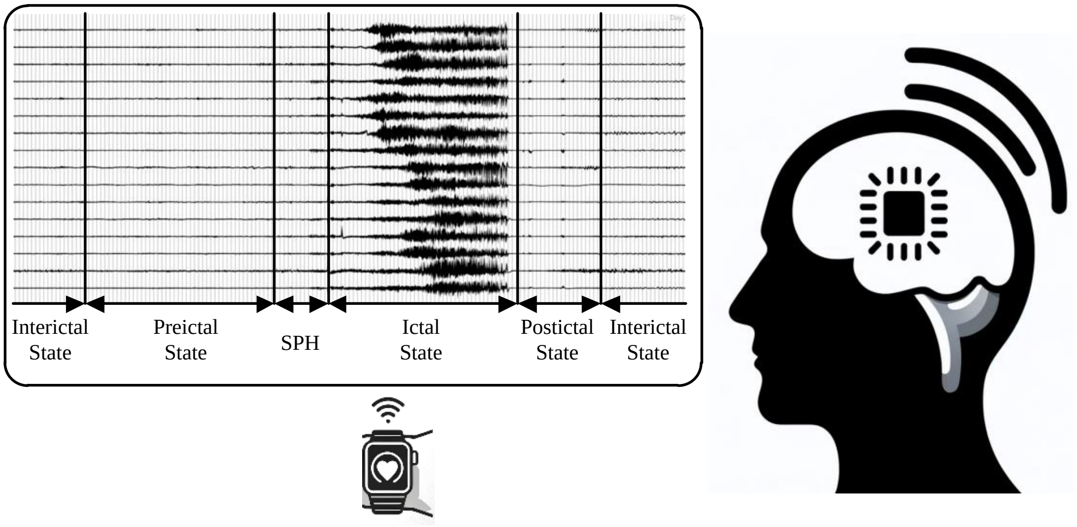


Figure 6.1 A conceptual framework for epilepsy seizure prediction in wearable biomedical devices

Table 6.1 Per-subject dataset characteristics: number of channels, sampling rates, and memory footprints before and after compression

Subject	**Channels**	**Before compression**		**After compression**	
		Sampling rate (Hz)	**Memory footprint (GB)**	**Sampling rate (Hz)**	**Memory footprint (MB)**
Dog 1	16	400	3.9	400	44.3
Dog 2	16	400	4.2	400	47.6
Dog 3	16	400	11.6	400	132.0
Dog 4	16	400	6.9	400	79.1
Dog 5	15	400	3.5	400	39.5
Patient 1	15	5000	6.9	400	5.6
Patient 2	24	5000	8.8	400	7.9

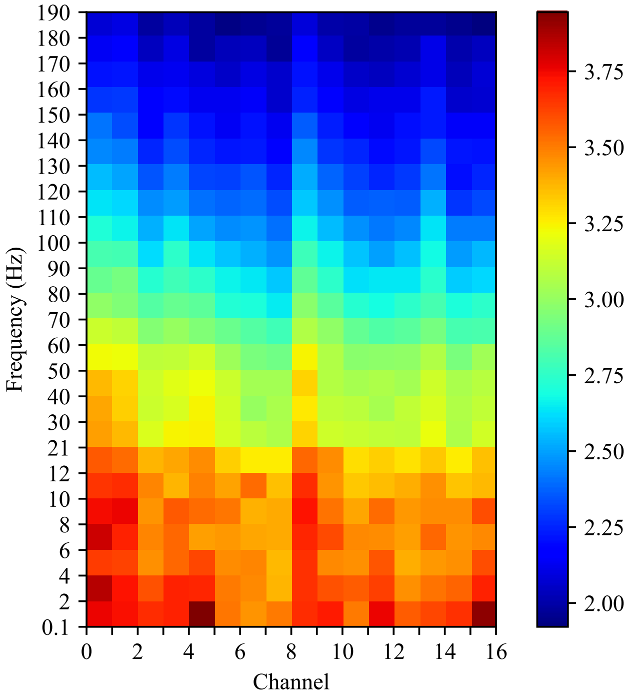


Figure 6.2 A 20-second 16-channel compressed data frame of Dog 1

The remainder of this paper includes the description of the adopted methodology in Section 6.3, the analysis of the results in Section 6.4, and finally the conclusion in Section 6.5.

6.3 Methodology

6.3.1 Dataset and Compression

In this work, the widely used AES dataset serves as a benchmark to compare various seizure prediction methodologies [32]. The dataset contains iEEG recordings from five canines and two human patients with epilepsy. Details about the dataset are presented in Table 6.1. The compression phase involves data segmentation, resampling, and a STFT. Each data frame includes a 20-second segment of the preictal or interictal state. Then, each data frame is resampled at 400 Hz and processed through a fifth-order Butterworth band-pass filter ranging from 0.1 Hz to 190 Hz, to facilitate subsequent STFT processing. STFT converts 20-second multi-channel iEEG signals into a frequency-channel representation across 24 frequency bands, as previously described in our work [87]. The compressed data frame for feeding into the CNN is shown in Figure 6.2.

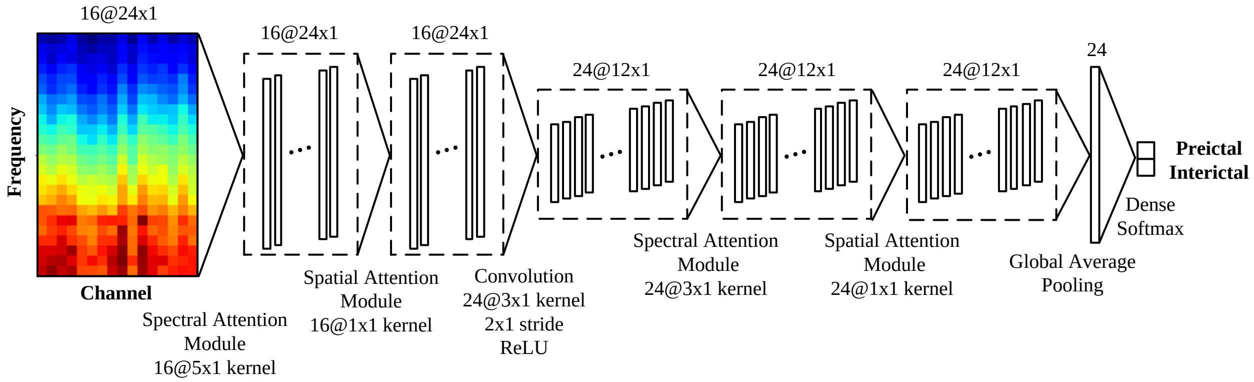

16@24x1
Frequency
Channel
Spectral Attention Module 16@5x1 kernel
16@24x1
Spatial Attention Module 16@1x1 kernel
16@24x1
Convolution 24@3x1 kernel 2x1 stride ReLU
24@12x1
Spectral Attention Module 24@3x1 kernel
24@12x1
Spatial Attention Module 24@1x1 kernel
24@12x1
Global Average Pooling
24
Dense Softmax
Preictal
Interictal

Figure 6.3 The overall architecture of the proposed attention-driven CNN architecture

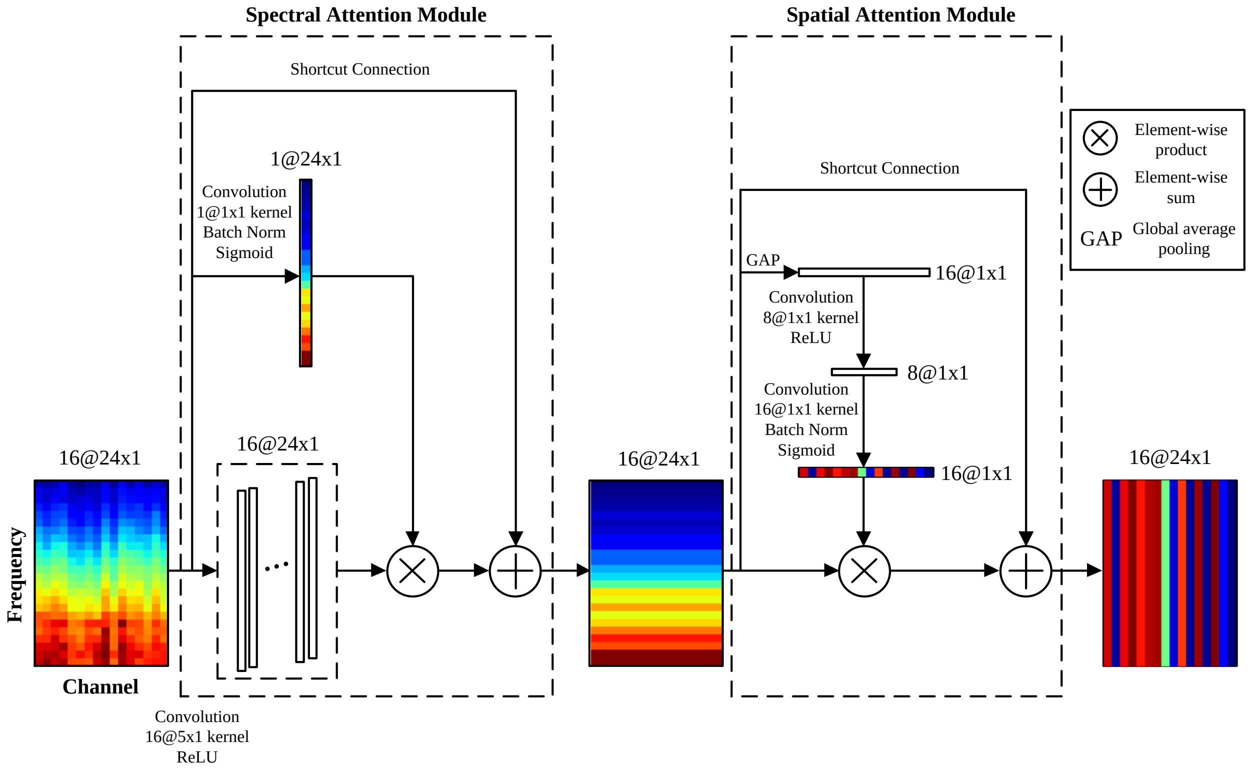

Spectral Attention Module
Shortcut Connection
1@24x1
Convolution 1@1x1 kernel Batch Norm Sigmoid
16@24x1
16@24x1
Frequency
Channel
Convolution 16@5x1 kernel ReLU
16@24x1
Spatial Attention Module
Shortcut Connection
GAP
16@1x1
Convolution 8@1x1 kernel ReLU
8@1x1
Convolution 16@1x1 kernel Batch Norm Sigmoid
16@1x1
16@24x1
Element-wise product
Element-wise sum
GAP Global average pooling

Figure 6.4 The basic structure of the spectral-spatial attention block

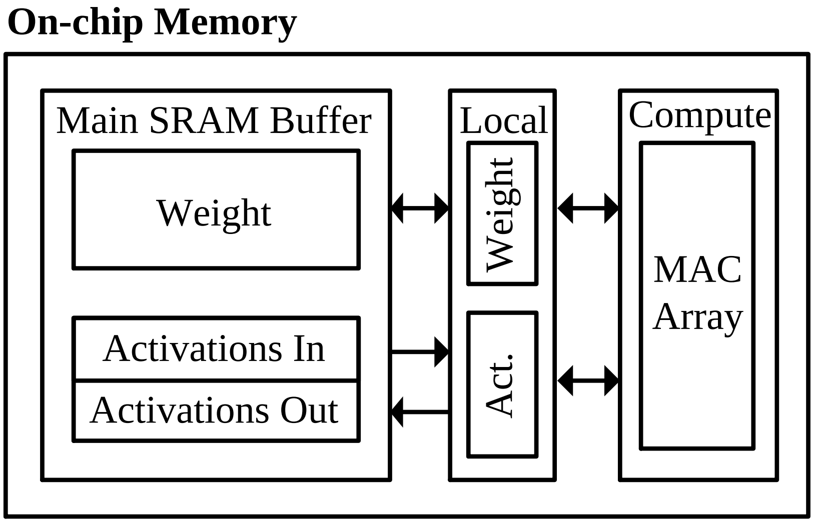


Figure 6.5 High-level overview of energy model

6.3.2 Attention-Driven CNN Architecture

Recent advances in bioinformatics have highlighted the effectiveness of the lightweight 1D CNN in the analysis of EEG signals [98–100], which merges temporal and spatial information within its local receptive field to create enhanced feature representations. In this work, inspired by the attention mechanism [101] and 1DSCNN [87], a compact $S^3$1DCNN is proposed to improve the interpretable ability to analyze non-stationary EEG signals by dynamically weighting the significance of various features. The overall architecture is shown in Figure 6.3, and a spectral-spatial attention block is shown in Figure 6.4. As depicted in Figure 6.3, first, 16-channel EEG signals undergo processing through a spectral-spatial attention block to highlight dynamic information interacting between space and frequency. Subsequently, a single 1D CNN block extracts more complex and abstract features. Then, another spectral-spatial attention block is employed to augment the model's capacity for generalization further. As a final step, a global average pooling follows to downsample all the features, which is then directly fed into a Softmax layer to perform classification. In Figure 6.4, a spectral-spatial attention block sequentially consists of a spectral attention module and a spatial attention module. These two modules play a crucial role in identifying which frequencies and channels should be emphasized or suppressed, thereby facilitating the computation of complementary attention.

6.3.3 Training Configuration

The labeled dataset is randomly partitioned for each subject into an 80% training set and a 20% test set through a five-fold stratified cross-validation [95]. Next, 20% of the training set is designated as the validation set to mitigate overfitting. Due to the highly skewed distribution between the preictal and interictal states [77], class weights inversely proportional to the counts of preictal and interictal states are applied to the loss function during the training phase. The Adam optimizer is used for training, with β_1 of 0.9, β_2 of 0.999, and an initial learning rate of 0.001. The learning rate is configured with an exponential decay schedule, featuring a decay rate of 0.99 and decay steps set at 500. The training phase ends when there is no further improvement in validation accuracy. The model of this work is implemented in Python 3.9, utilizing Keras 2.10.0 with a Tensorflow 2.10.1 backend on a single NVIDIA Tesla V100 GPU.

6.4 Results and Analysis

6.4.1 Evaluation Metrics

Statistical Assessment

Sensitivity, FPR, and AUC are employed to evaluate and benchmark our method against recent state-of-the-art methods. Sensitivity is calculated as the proportion of 20-second seizure segments correctly identified out of the total number of 20-second seizure segments. FPR is defined as the number of false positives occurring per hour [34]. AUC evaluates the performance of a classifier.

Energy Assessment

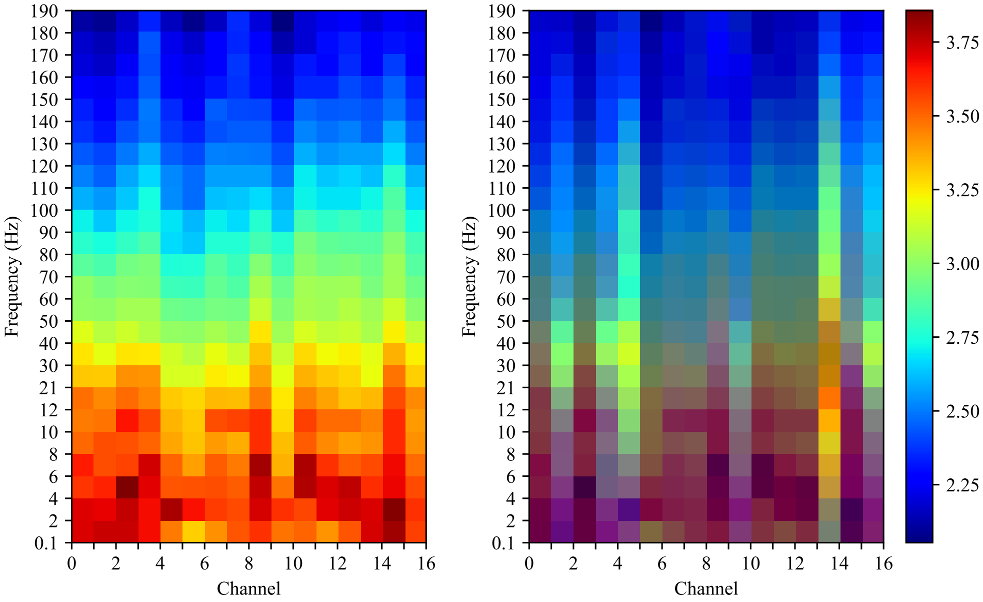
This work focuses on forecasting epileptic seizures using low-power wearable devices. A simplified, parameterized energy model grounded in the framework of [71] is applied to estimate energy consumption, which leverages a 28nm FDSOI technology. The energy consumption is estimated only on the assumption of on-chip memory access. This process initially loads weights and activations into the SRAM buffer. After that, they are directly fed into the MAC array through the local buffer for calculation. Thus, the energy consumption for inference comprises the compute energy, on-chip weight access energy, and activation access energy, as illustrated in Figure 6.5.

Table 6.2 Per-subject statistical assessment results for the AES dataset: sensitivity, FPR, and AUC

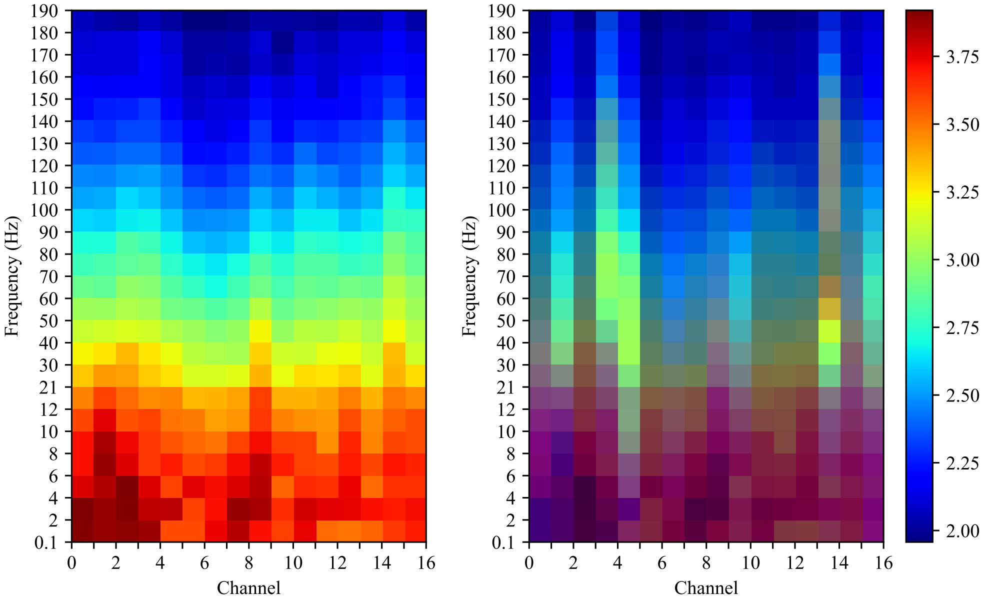
Subject	**Sensitivity (%)**	**FPR (/h)**	**AUC**
Dog 1	85.3±1.5	0.008±0.001	0.937±0.005
Dog 2	97.1±0.9	0.001±0.000	0.998±0.001
Dog 3	89.4±1.6	0.001±0.000	0.984±0.002
Dog 4	88.7±1.2	0.002±0.000	0.972±0.002
Dog 5	97.0±1.9	0.001±0.000	0.999±0.000
Patient 1	97.6±1.9	0.008±0.004	0.998±0.001
Patient 2	90.0±5.6	0.035±0.011	0.976±0.006
Average	**92.1±2.1**	**0.008±0.002**	**0.980±0.003**

Table 6.3 Comparison with other state-of-the-art methods

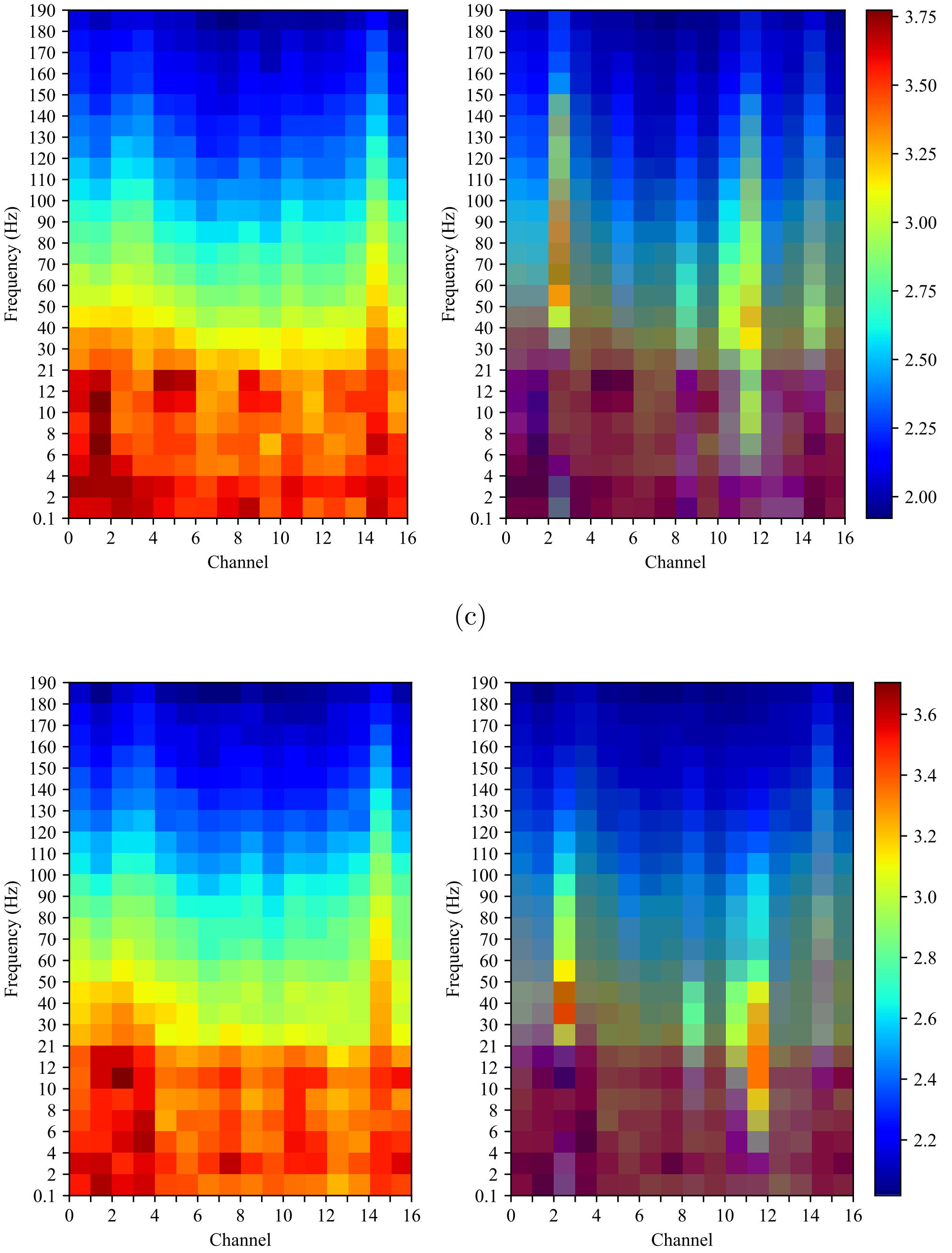
Method	**Sensitivity (%)**	**FPR (/h)**	**AUC**	**Model size**	**Energy (μJ)**
Brinkmann et al. [65]	-	-	0.860	-	-
Korshunova et al. [38]	-	-	0.810	0.56 MB	1.24
Truong et al. [34]	75.0±0.0	0.210±0.040	-	0.76 MB	-
Liu et al. [61]	-	-	0.837	1.79 MB	9.43
Zhao et al. [19]	93.48	0.063	0.977	44.52 kB	153.82
This work	**92.1±2.1**	**0.008±0.002**	**0.980±0.003**	**21.50 kB**	**0.21**


190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
21
12
10
8
6
4
2
0.1
Frequency (Hz)
0
2
4
6
8
10
12
14
16
Channel
3.75
3.50
3.25
3.00
2.75
2.50
2.25

(a)


190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
21
12
10
8
6
4
2
0.1
Frequency (Hz)
0
2
4
6
8
10
12
14
16
Channel
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00

(b)



(c)

(d)

Figure 6.6 Original and Grad-CAM data frames of interical and preictal iEEGs for Dog 2: (a)-(b) two examples of true negative results for interictal prediction; (c)-(d) two examples of true positive results of preictal prediction

6.4.2 Performance Analysis

Table 6.2 presents the statistical evaluation of the proposed $S^3$1DCNN model through a five-fold stratified cross-validation, demonstrating an average sensitivity of 92.1%, an average FPR of 0.008/h, and an average AUC of 0.980 across all subjects in the dataset. Table 6.3 compares our method against other state-of-the-art methods using the same dataset. The model size is reported in terms of 32-bit floating-point precision parameters. The comparison results show that the proposed $S^3$1DCNN model outperforms existing state-of-the-art methods regarding AUC and model size while maintaining the lowest energy consumption at 0.21 μJ per inference. It demonstrates our method's promising application potential in low-power and energy-efficient wearable devices.

6.4.3 Interpretability Analysis

Grad-CAM [57] is applied to identify and visualize the areas of the input data frame that contribute most to the model's output. Figure 6.6 shows original and Grad-CAM data frames of interical and preictal iEEGs for Dog 2. It should be noted that Grad-CAM analysis demonstrates distinct emphases on channels (ch2, ch4, ch5, ch10, ch14, and ch16 for interictal state, ch3, ch9, ch11, ch12 and ch15 for preictal state) and frequency ranges between preictal and interictal data frames, which presents our model is potential to enhance understanding of functional brain connectivity and to localize regions of epilepsy onset accurately.

6.5 Conclusions

This work proposed an attention-driven $S^3$1DCNN for predicting epileptic seizures. It can interpretably analyze non-stationary EEG signals while consuming very little energy, as required with wearable biomedical devices. The proposed $S^3$1DCNN outperforms recently reported state-of-the-art methods in terms of AUC, model size, and energy consumption on the AES dataset. This method shows excellent potential for hardware-friendly low-power biomedical wearable devices.

CHAPTER 7 SUPPLEMENTARY RESULTS

7.1 Extending the Findings of Chapter 4

In Chapter 4, a preliminary quantization approach is employed to evaluate the impact of various bit widths on model performance. To illustrate the difference from the quantization method proposed in Chapter 5, Table 7.1 provides detailed results on AUC scores and model size across various quantization levels. Compared with Table 5.3 in Chapter 5, the quantization method proposed in Chapter 5 achieves better performance due to several improvements such as the incorporation of a scaling factor, heterogeneous quantization granularity, and better training strategy. To ensure low latency and minimal quantization error, per-tensor quantization is chosen for activation values, while per-channel quantization is selected for weights in Chapter 5. A more detailed comparison and discussion will be conducted in Section 7.2.2.

Table 7.1 Full precision and various quantization bit widths results for the AES dataset: average AUC, precision loss, model size, and compression factor

	Full Precision	16-bit Quant.	8-bit Quant.	4-bit Quant.	2-bit Quant.	1-bit Quant.
Average AUC	0.979	0.976	0.976	0.974	0.956	0.890
Precision loss (%)	0.00	0.31	0.31	0.51	2.35	9.09
Model size (kB)	21.32	10.86	5.62	3.01	1.70	1.05
Compression factor	1.00x	1.96x	3.79x	7.08x	12.54x	20.30x

7.2 Further Analysis on the Seizure Prediction Method Reported in Chapter 6

7.2.1 Extending the Findings of Interpretability

In Chapter 6, Grad-CAM is used to visualize and interpret decision-making processes with the proposed $S^3$1DCNN. The methodology and principle of Grad-CAM is illustrated in Section 3.3.2. Building on this groundwork, supplementary results in Figure 7.1 and Figure 7.2

are presented. These results provide a more comprehensive understanding of the effectiveness and generalization of Grad-CAM in offering insights into the reasoning mechanisms of the proposed $S^3$1DCNN, thereby enhancing the transparency and reliability of seizure prediction.

On the one hand, Figure 7.1 presents six randomly selected examples from the test set, comparing the original data frames with the Grad-CAM outputs. These examples show true negative predictions for the interictal period. The left side of Figure 7.1a, Figure 7.1b, Figure 7.1c, Figure 7.1d, Figure 7.1e and Figure 7.1f displays the original data, while the right side shows the corresponding Grad-CAM output. The mask of Grad-CAM output highlights which channels and frequency components have the most significant impact on prediction results. This approach offers a new perspective for investigating seizure onset zones and mechanisms of ictogenesis.

In general, from the highlighted parts of the Grad-CAM output on the right side of Figure 7.1a, Figure 7.1b, Figure 7.1c, Figure 7.1d, Figure 7.1e and Figure 7.1f, we can observe that specific channels and frequency components are consistently emphasized. It is important to clarify that, for example, in Figure 7.1a, the horizontal axis represents channels, while the vertical axis shows frequency ranges. The channel between numbers 0 and 1 is channel 1, the channel between numbers 1 and 2 is channel 2, and so on. The heat maps in Figure 7.1 highlight that channels 2, 4, 5, 10, 14, and 16 contribute more significantly to generating true negative results in interictal prediction. For instance, channel 7 also exhibits more activity during interictal states in Figure 7.1c, Figure 7.1d, Figure 7.1e and Figure 7.1f, which shows that channel 7 may have exhibited activity due to its sensitivity to subtle patterns or fluctuations in neural signals during interictal states. In addition, channel 9 exhibits only minimal activity in Figure 7.1e. Although channel 9 may not contribute significantly to generating true negative results, its minimal activity suggests it could detect underlying neural dynamics or noise unrelated to the specific predictive features targeted in interictal states.

From a frequency perspective, channels 5 and 16 in Figure 7.1 capture a broad spectrum of neural activity, ranging from 0.1 to 180 Hz. This includes low-frequency signals that may reflect slow cortical potentials and baseline brain activity, as well as high-frequency signals, often linked to rapid neural processes such as action potentials and fast oscillations. Channels 2, 4, 14 in Figure 7.1 except Figure 7.1a capture gamma frequencies [65], ranging from 30 to 180 Hz. These high-frequency oscillations are often associated with high-level brain functions and can reflect synchronous neural activity during various brain states, which can be crucial for accurate interictal prediction.

On the other hand, Figure 7.2 presents six randomly selected examples from the test set, comparing the original data frames with the Grad-CAM outputs. These examples show

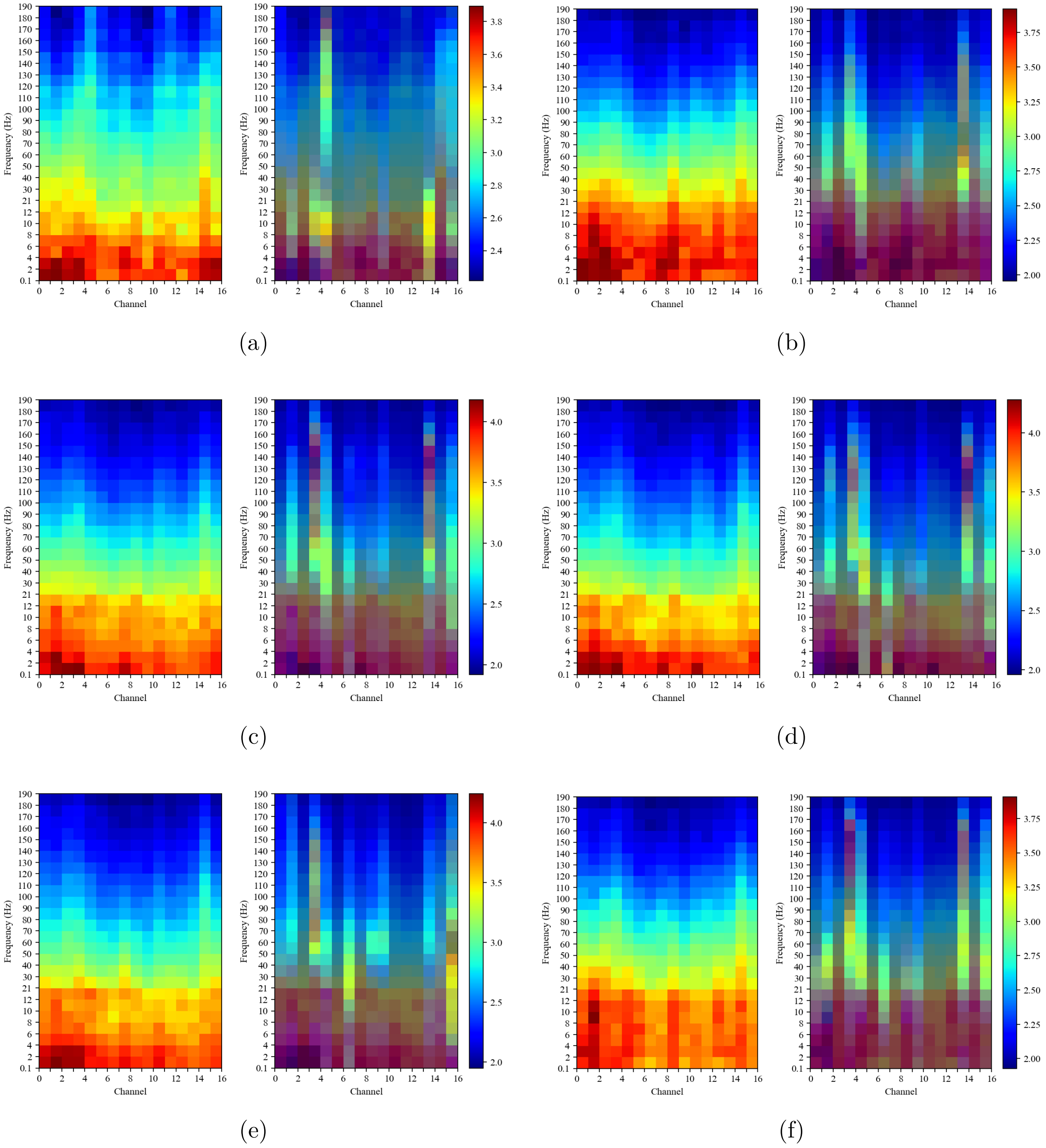
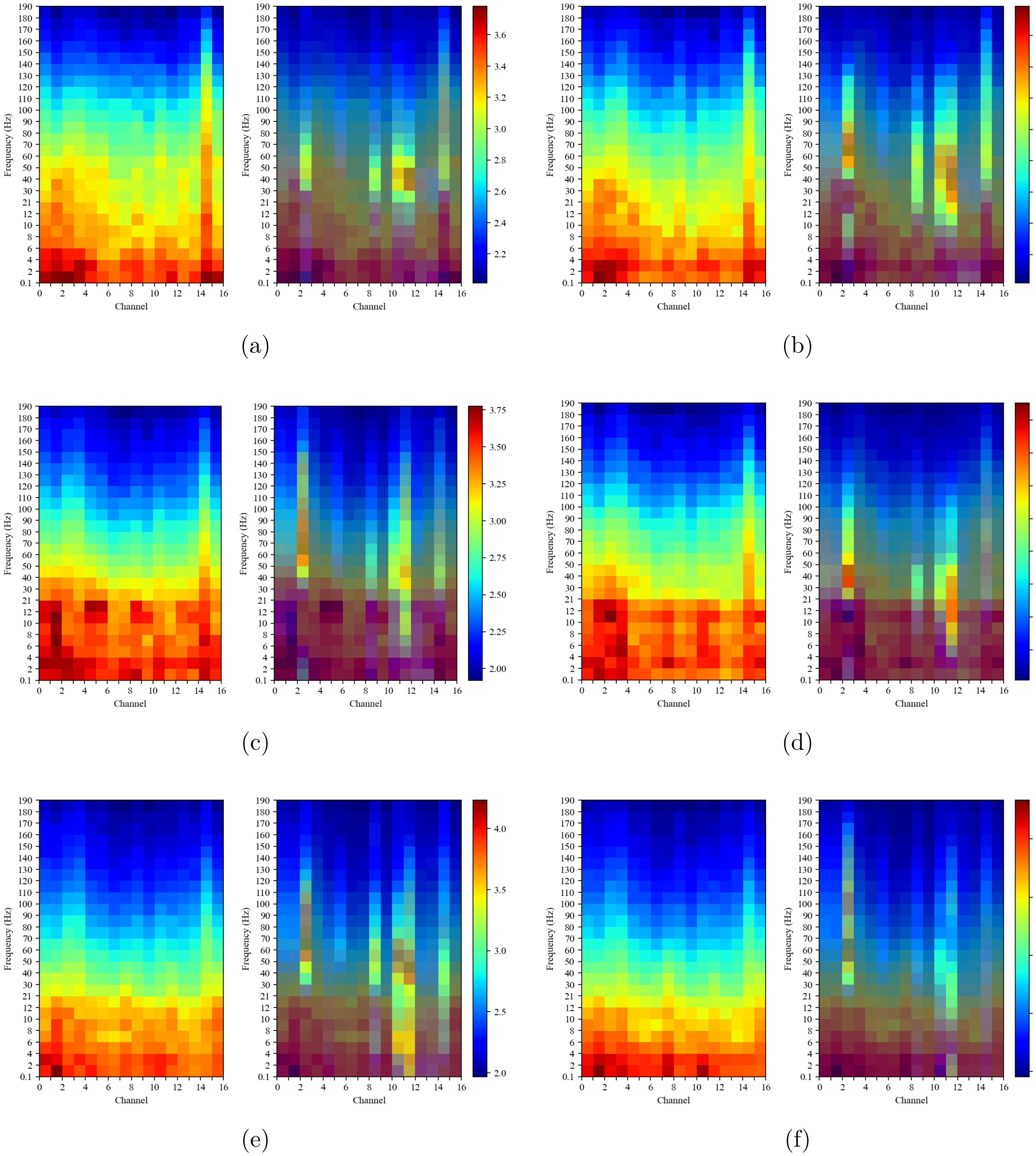


Figure 7.1 (a)-(f) present six randomly selected test set examples, comparing the original data frames with Grad-CAM outputs, which show true negative predictions for the interictal period



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7.2 (a)-(f) present six randomly selected test set examples, comparing the original data frames with Grad-CAM outputs, which show true positive predictions for the preictal period true positive predictions for the preictal period. The left side of Figure 7.2a, Figure 7.2b, Figure 7.2c, Figure 7.2d, Figure 7.2e and Figure 7.2f displays the original data, while the right side shows the corresponding Grad-CAM output. Likewise, the mask of Grad-CAM output highlights which channels and frequency components have the most significant impact on prediction results.

In contrast to Figure 7.1, in general, Figure 7.2 highlights that channels 3, 9, 11, 12, and 15 play a more significant role in producing true positive results in preictal prediction. These channels differ markedly from those used to predict interictal status. Notably, channel 6 exhibits increased activity during preictal states in Figure 7.2c and Figure 7.2e. This suggests that channel 6 may be sensitive to subtle patterns or fluctuations in neural signals during preictal states.

From a frequency perspective, all highlighted channels 3, 6, 9, 11, 12 and 15 in Figure 7.2 cover the low-gamma band [65], ranging from 30 to 70 Hz. These gamma oscillation channels capture critical features relevant to the epileptiform synchronization of interneurons for preictal prediction, linking the seizure onset zone to the surrounding epileptogenic zone [102].

To this end, Figure 7.1 indicates that channels 2, 4, 5, 10, 14 and 16 exhibit more activity during interictal states. In contrast, channels 3, 9, 11, 12 and 15 demonstrate increased activity during preictal states, as shown in Figure 7.2. This distinction indicates that different regions of the brain and their corresponding activities are crucial in various phases of the seizure cycle. Meanwhile, the analysis shows that the gamma frequency band across highlighted channels is essential for distinguishing between preictal and interictal states. Gamma oscillations play a key role in preictal seizure prediction.

7.2.2 Extending the Findings of Compatibility

As mentioned before, the 1DSCNN was first proposed to predict epileptic seizures and validated on the AES dataset in Chapter 4. Then, a preliminary quantization approach was employed to assess how different bit widths affect model performance. In Chapter 5, the proposed 1DSCNN developed in Chapter 4 is further evaluated to show its ability to generalize on the AES and CHB-MIT datasets. In contrast to Chapter 4, which only assesses model size, Chapter 5 presents a comprehensive on-chip energy model that evaluates energy costs during the inference phase, including computing energy, weight access energy, and activation access energy, as detailed in Section 5.3.3. Subsequently, a well-refined fixed-precision quantization method has been applied to the two datasets for high energy efficiency with only minor reductions in performance. It achieves better results than those detailed in Chapter 4, thanks to improvements in scaling factor, quantization granularity, and training strategy.

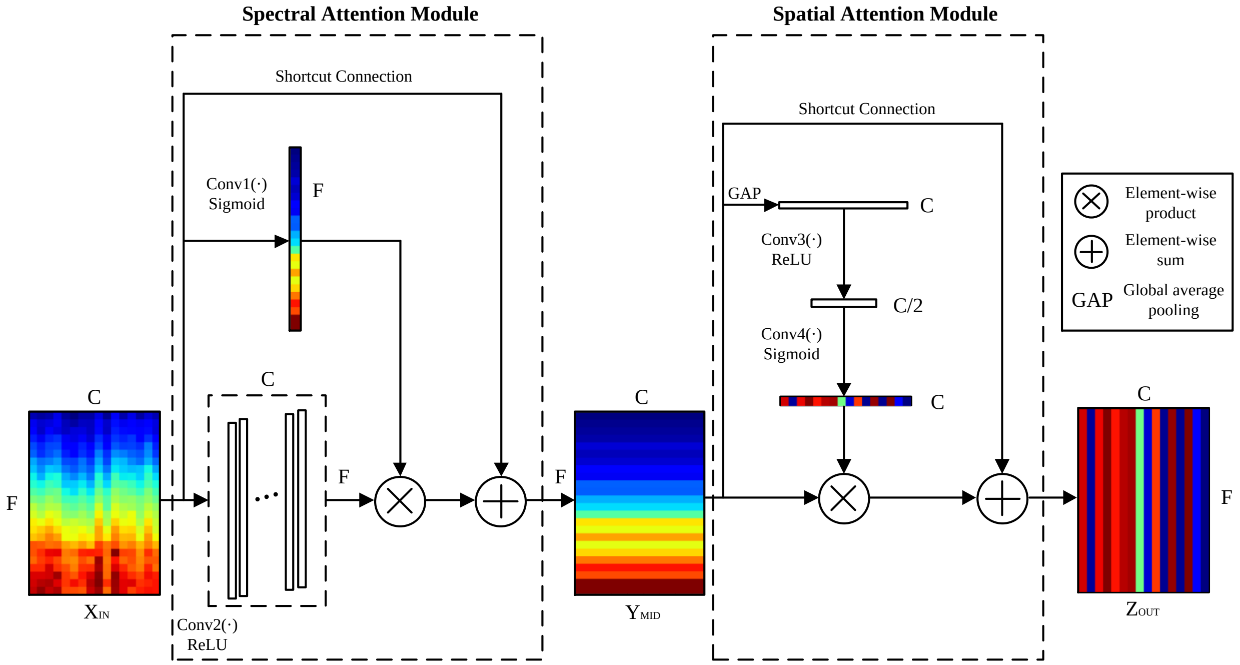


Figure 7.3 The computational flow of the spectral-spatial attention module

Specifically, in Chapter 4, the scaling factors for the weights, biases, and activations were determined per layer, while in Chapter 5, the scaling factors for the biases and activations remain per layer. Still, the scaling factors for the weights are adjusted per channel. This adjustment can reduce quantization errors, including rounding and clipping errors, which positively impacts the performance of higher bit-width quantization (such as 16-bit and 8-bit quantization). Regarding training strategy, both chapters use the same strategy to train the floating-point baseline model. For the quantization model, Chapter 4 follows the same training strategy as the floating-point baseline. This involves using the Adam optimizer over 100 epochs, with β_1 of 0.9, β_2 of 0.999, and an initial learning rate of 0.001, which is halved every 50 epochs. However, in Chapter 5, the Adam optimizer is also used for training with the same β_1 and β_2 values, but over 200 epochs, with the learning rate starting at 0.001 and halving every 50 epochs. This is because the quantization model requires more time and fine-tuning of the learning rate to achieve optimal performance. Finally, a low-complexity mixed-precision search method is also proposed to improve performance for hard-to-predict seizures.

In Chapter 6, inspired by the attention mechanism and 1DSCNN, an attention-driven $S^3$1DCNN is introduced to precisely forecasting epileptic seizures, showcasing its capability to interpret and analyze spatio-temporal, non-stationary iEEG recordings for accurate identification of epilepsy onset regions. The main difference between this chapter and Chapter 4 and 5 is that the convolutional layer in Figure 5.3 is replaced by the spectral-spatial attention module. The computational flow of the spectral-spatial attention module is shown in Figure 7.3. Assume that an input feature map X_{IN} is expressed as $X_{IN} = [x_1, x_2, \dots, x_F]$, where each $x_i \in \mathbb{R}^{1\times C}$ and $i = 1, 2, \dots, F$. According to Section 3.3.1, the intermediate output Y_{MID} can be described as:

$$Y_{MID} = X_{IN} + \sigma(Conv1(X_{IN})) \cdot \delta(Conv2(X_{IN}))\,, \tag{7.1}$$

where $Y_{MID} = [y_1, y_2, \dots, y_C]$, where $y_j \in \mathbb{R}^{F\times 1}$. σ is the Sigmoid function and δ is the ReLU function. $Conv1$ and $Conv2$ represent convolutional operations that encode feature information across different channels. Then, following the same approach outlined in Section 3.3.1, the final output Z_{OUT} can be expressed as:

$$\begin{aligned} Z_{OUT} = Y_{MID} + [&y_1\sigma(Conv4(\delta(Conv3(\frac{1}{1\times F}\sum_{k=1}^{F} y_1(k))))), \\ &y_2\sigma(Conv4(\delta(Conv3(\frac{1}{1\times F}\sum_{k=1}^{F} y_2(k))))), \dots, \\ &y_C\sigma(Conv4(\delta(Conv3(\frac{1}{1\times F}\sum_{k=1}^{F} y_C(k)))))]\,, \end{aligned} \tag{7.2}$$

where σ is the Sigmoid function, and δ is the ReLU function. $Conv3$ and $Conv4$ represent convolutional operations designed to capture dependencies among channels, both using a kernel size of 1. Thus, it can be seen that this spectral-spatial attention module can merge spectral and spatial information within its local receptive field to create enhanced feature representations by dynamically weighting the significance of various features. Meanwhile, this attention module can integrate effectively with Grad-CAM to determine which channels and frequency components have the most significant impact on prediction results, which can have the potential to clarify the location of seizure onset. In addition, the on-chip energy model presented in Chapter 5 is also used to evaluate the energy costs of this model.

For a fair comparison, each baseline model in this thesis for the AES dataset is analyzed to demonstrate the costs and benefits, as shown in Table 7.2. Compared to Chapter 4, Chapter 5 only adds energy consumption as an additional evaluation factor for the baseline model. Thus, the main comparison focuses on the methods presented in Chapter 5 and Chapter 6. The $S^3$1DCNN proposed in Chapter 6 achieves comparable sensitivity and FPR to those in the other chapters, but demonstrates a slightly better AUC. Meanwhile, the $S^3$1DCNN in Chapter 6 has a 0.84% larger model size and 5.1% higher energy consumption

Table 7.2 Methods comparison of each chapter for the AES dataset

Method	**Dataset**	**# of subjects**	**Classifier**	**Sensitivity (%)**	**FPR (/h)**	**AUC**	**Model size**	**Energy (μJ)**	**Preictal duration**	**Interpretable**
Chapter 4	AES	5 dogs, 2 patients	1DSCNN	94.44	0.011	0.979	21.32 kB	-	60 min	No
Chapter 5	AES	5 dogs, 2 patients	1DSCNN	94.44	0.011	0.979	21.32 kB	0.198	60 min	No
Chapter 6	AES	5 dogs, 2 patients	$S^3$1DCNN	92.1±2.1	0.008±0.002	0.980±0.003	21.50 kB	0.208	60 min	Yes

than the methods in other chapters. Still, it offers interpretability to clarify the seizure onset zone. This feature can provide healthcare professionals with valuable insights into seizure onset regions for better preventive measures and medication interventions. Moreover, combining the attention mechanism with Grad-CAM to identify important parts of the input contributing to the model's predictions is the first approach applied for epileptic prediction in Chapter 6. To this end, in contrast to the methods in Chapter 4 and Chapter 5, the attention-driven $S^3$1DCNN introduced in this chapter results in a slight increase in model size and energy consumption. Yet, it marginally outperforms them and offers interpretability to clarify the location of the seizure onset.

CHAPTER 8 GENERAL DISCUSSION

This chapter begins by revisiting the contributions of this thesis in addressing the research objectives outlined in Chapter 1. Then, it goes on to interpret and compare the results. The chapter concludes by exploring the broader implications and significance of the conducted research and the proposed methods.

8.1 Addressing Research Objectives

This thesis focuses on developing a robust epileptic seizure prediction engine specifically designed for wearable medical devices. The primary purpose of this initiative is to discover and implement practical approaches that can quickly provide information on seizure prediction to patients with epilepsy, improving their ability to manage and anticipate seizures. This work aims to develop high-performance, energy-efficient, and interpretable methods for seizure prediction in resource-limited wearable devices that can potentially be suitable for daily use. One of the principal challenges addressed in this work is the inherent trade-off between the performance of the seizure prediction algorithm and the device's energy consumption, which is a critical issue in designing wearable or implantable medical devices with limited resources.

To systematically tackle the challenge of our research aim, the core aim of this thesis has been structured around three different objectives, which include:

- The first objective was to design a high-performance, patient-specific CNN with a tiny model size for effective seizure prediction, which helps alleviate patient anxiety (low FPR) and enable preventive measures or medication interventions (high sensitivity). To achieve the first objective, a 1DSCNN is proposed to predict epilepsy seizures. This model is designed with a very competitive small size that is suitable for wearable biomedical devices. Compared to state-of-the-art methods, the proposed 1DSCNN achieves the best performance with the smallest model size on the AES dataset. In addition, a preliminary quantization scheme is also applied to evaluate the impact of various bit widths on model performance, facilitating its deployment in wearable biomedical devices.

- The second objective was to further explore an energy-efficient, patient-specific DL algorithm for effectively predicting epileptic seizures, addressing the trade-off between performance and energy consumption in resource-limited wearable medical devices. In contrast to the first objective, which only assesses model size, a comprehensive on-chip energy model is presented to evaluate energy consumption to achieve the second objective. Then, our proposed 1DSCNN in the first objective was further evaluated to demonstrate its generalization on the AES and CHB-MIT datasets, which surpasses state-of-the-art methods in terms of sensitivity, FPR, AUC, model size and energy consumption. Next, a hardware-friendly, fixed-precision quantization scheme was implemented and tested on the two datasets, resulting in excellent energy efficiency with minimal performance losses. Finally, a near-optimal, low-complexity mixed-precision search method is also suggested, which has the potential to improve model performance for subjects with epileptic seizures that are challenging to predict.

- The third objective was to develop a high-performance, interpretable CNN with low energy consumption for accurately forecasting epileptic seizures, which is intended to be transparent and understandable to healthcare professionals. To achieve the third objective, an attention-driven $S^3$1DCNN was proposed to forecast seizures, featuring its interpretable ability to analyze spatiotemporal non-stationary iEEG recordings for precisely localizing regions of epilepsy onset. In addition, the proposed $S^3$1DCNN outperforms state-of-the-art methods regarding AUC, model size, and energy consumption on the AES dataset, which shows its excellent potential for low-power biomedical wearable devices.

To this end, the core aim of this thesis can be summarized in a short form as the following three objectives: enhancing the performance and responsiveness of the seizure prediction algorithm with small model size, optimizing the energy efficiency of the seizure prediction system to prolong device usability, and improving the interpretability of the predictive results to ensure that healthcare professionals can easily understand the information provided. Each objective is designed step by step to significantly contribute to developing an optimal balance between effectiveness and practicality, pushing the boundaries of what is currently possible in wearable epilepsy management technologies.

8.2 Interpretation of Results

After reviewing the research objectives of this thesis, the results of each objective are analyzed and compared below to provide a better understanding of our contributions and how they relate to each other:

- For the first objective, the proposed 1DSCNN outperforms recent state-of-the-art methods, achieving an average sensitivity of 94.44%, an average FPR of 0.011 per hour, and an average AUC of 0.979 on the AES dataset. The model is notably tiny, with a size of only 21.32 kB. In addition, a preliminary quantization scheme was applied to the proposed 1DSCNN model for evaluating the impact of different bit widths on model performance. After adapting the model to 8-bit quantization, its size was reduced by 3.79 times with only a minimal loss of 0.31% in AUC score precision. The results demonstrate that the proposed method obtains the best performance compared to recent state-of-the-art methods, which showcases the promising potential to help alleviate patient anxiety (low FPR) and enable preventive measures or medication interventions (high sensitivity).

- For the second objective, firstly, our proposed 1DSCNN based on average-pooling STFT is further evaluated to demonstrate its ability to generalize on the AES and CHB-MIT datasets, which achieves sensitivities of 94.44%, 96.14%, FPR of 0.011/h, 0.018/h, and AUC of 0.979, 0.996, for model sizes of 21.32 kB, 23.62 kB, with estimated energy consumption of 0.20 μJ, 0.28 μJ when processing data from the AES and CHB-MIT datasets, respectively. Then, a refined fixed-precision quantization scheme was proposed, implemented, and characterized on the two considered datasets. Notably, compared to the baseline mentioned above, 8-bit quantization reduces the model size by 3.79 times from 21.32 kB to 5.62 kB with only a 0.20% average AUC score loss for the AES dataset, while 4-bit quantization reduces the model size by 5.95 times from 21.32 kB to 3.01 kB with the same 0.20% average AUC score loss for the CHB-MIT dataset. This quantization scheme reduces energy consumption by representing model parameters with fewer bits, reducing model size and memory usage. This is especially critical for wearable biomedical or implanted devices with limited resources. To this end, the AES dataset is more sensitive to quantization errors (rounding and clipping errors) than the CHB-MIT dataset. Additionally, the 8-bit quantization results surpass those of the first objective due to refined factors, including the scaling factor, quantization granularity, and training strategy. Finally, a low-complexity mixed-precision search method is proposed to improve model performance for subjects with hard-to-predict epileptic seizures, which achieves a model size reduction of 7.06x, with only 0.20% AUC score loss in the AES dataset, and a size reduction of 8.59x, with only 0.30% AUC score loss in the CHB-MIT dataset. The results demonstrate that the proposed methods provide better performance than recent state-of-the-art methods, which address the trade-off between performance and energy consumption in resource-limited wearable medical devices.

- For the third objective, an attention-driven $S^3$1DCNN is proposed to forecast seizures accurately. The results demonstrate that the proposed method outperforms recent state-of-the-art methods, achieving an average sensitivity of 92.1%, an average FPR of 0.008/h, an average AUC of 0.980, and an estimated energy consumption of 0.21 μJ per inference on the AES dataset. Meanwhile, the proposed $S^3$1DCNN can analyze spatio-temporal non-stationary iEEG recordings for precisely localizing regions of epilepsy onset due to its interpretable ability. The results demonstrate that the proposed method effectively predicts seizures with minimal energy consumption and provides healthcare professionals with valuable insights into seizure onset regions for better preventive measures and medication interventions. This shows its excellent potential for low-power and energy-efficient biomedical wearable devices.

8.3 Comparison of Results

After interpreting the results related to each research objective, a comparison is conducted to illustrate the differences between the outcomes of various objectives and to benchmark them against state-of-the-art methods. Table 8.1 compares the results from each chapter in terms of sensitivity, FPR, AUC, model size, energy consumption, preictal duration and interpretibility, with previously reported leading epileptic seizure prediction methods for the AES dataset. As a supplement to validate generalization in Chapter 5, Table 8.2 compares the results of the second research objective in terms of sensitivity, FPR, AUC, model size, energy consumption, preictal duration, with the recent state-of-the-art methods for the CHB-MIT dataset.

All methods in Table 8.1 employ CNNs for classification, and the majority of methods [19, 38, 61] adopt a preictal duration of 60 minutes. In general, the 32-bit floating-point baseline model (1DSCNN) in Chapter 4, the 32-bit floating-point baseline model (1DSCNN) and the 16-bit quantization model in Chapter 5, the model in Chapter 6 outperform prior approaches [19, 34, 38, 61], achieving the best sensitivity, FPR, and AUC while maintaining the lowest model size and energy consumption. In particular, the 8-bit quantization model and the mixed-precision model in Chapter 5 achieve the same AUC of 0.977 as in [19], but with model sizes of 26% and 14%, and significantly lower energy consumption of both nearly one fifteen-thousandth.

Compared to Chapter 4, Chapter 5 validates the generalization ability of the proposed 1DSCNN on the AES and CHB-MIT datasets, provides a specific energy model to evaluate energy consumption, refines the fixed-precision quantization scheme to reduce quantization errors, and proposes a mixed-precision search method to improve model performance for

Table 8.1 Comparison to prior state-of-the-art methods for the AES dataset

Method	**Dataset**	**# of subjects**	**Classifier**	**Sensitivity (%)**	**FPR (/h)**	**AUC**	**Model size**	**Energy (μJ)**	**Preictal duration**	**Interpretable**
Korshunova et al. 2017 [38]	AES	5 dogs, 2 patients	CNN	-	-	0.810	0.56 MB	1.24	60 min	No
Liu et al. 2019 [61]	AES	5 dogs, 2 patients	Multi-view CNN	-	-	0.837	1.79 MB	9.43	60 min	No
Truong et al. 2018 [34]	AES	5 dogs, 2 patients	CNN	75	0.21	-	0.76 MB	-	30 min	No
Zhao et al. 2021 [19]	AES	5 dogs, 2 patients	CNN	93.48	0.063	0.977	44.52 kB	153.82	60 min	No
Chapter 4	AES	5 dogs, 2 patients	1DSCNN	94.44	0.011	0.979	21.32 kB	-	60 min	No
			16-bit Quant.	-	-	0.976	10.86 kB	-	60 min	
			8-bit Quant.	-	-	0.976	5.62 kB	-	60 min	
			4-bit Quant.	-	-	0.974	3.01 kB	-	60 min	
			2-bit Quant.	-	-	0.956	1.70 kB	-	60 min	
			1-bit Quant.	-	-	0.890	1.05 kB	-	60 min	
Chapter 5	AES	5 dogs, 2 patients	1DSCNN	94.44	0.011	0.979	21.32 kB	0.198	60 min	No
			16-bit Quant.	-	-	0.978	10.86 kB	0.123	60 min	
			8-bit Quant.	-	-	0.977	5.62 kB	0.099	60 min	
			4-bit Quant.	-	-	0.974	3.01 kB	0.092	60 min	
			2-bit Quant.	-	-	0.956	1.70 kB	0.089	60 min	
			1-bit Quant.	-	-	0.890	1.05 kB	0.088	60 min	
			MP Quant.	-	-	0.977	3.02 kB	0.094	60 min	
Chapter 6	AES	5 dogs, 2 patients	$S^3$1DCNN	92.1±2.1	0.008±0.002	0.980±0.003	21.50 kB	0.208	60 min	Yes

Table 8.2 Comparison to prior state-of-the-art methods for the CHB-MIT dataset

Method	**Dataset**	**# of subjects**	**Classifier**	**Sensitivity (%)**	**FPR (/h)**	**AUC**	**Model size**	**Energy (μJ)**	**Preictal duration**
Daoud et al. 2019 [43]	CHB-MIT	8 patients	Bi-LSTM	99.72	0.004	-	71.66 kB	-	60 min
Zhang et al. 2019 [77]	CHB-MIT	23 patients	CNN	92.2	0.12	0.90	-	-	30 min
Tang et al. 2020 [78]	CHB-MIT	24 patients	Multi-view CGRN	94.5	0.118	0.901	-	-	30 min
Wu et al. 2023 [75]	CHB-MIT	11 patients	CNN based ResNet	94.2	0.09	-	-	-	30 min
Truong et al. 2018 [34]	CHB-MIT	13 patients	CNN	81.2	0.16	-	0.77 MB	-	30 min
Zhao et al. 2021 [19]	CHB-MIT	10 patients	CNN	99.81	0.005	1	45.22 kB	141.46	30 min
Chapter 5	CHB-MIT	10 patients	1DSCNN	97.92	0.003	1	23.62 kB	0.277	30 min
		24 patients	1DSCNN	96.14	0.018	0.996	23.62 kB	0.277	30 min
			16-bit Quant.	-	-	0.994	12.39 kB	0.168	30 min
			8-bit Quant.	-	-	0.994	6.78 kB	0.134	30 min
			4-bit Quant.	-	-	0.994	3.97 kB	0.123	30 min
			2-bit Quant.	-	-	0.989	2.56 kB	0.119	30 min
			1-bit Quant.	-	-	0.967	1.86 kB	0.118	30 min
			MP Quant.	-	-	0.993	2.75 kB	0.122	30 min

subjects with hard-to-predict epileptic seizures.

Unlike the AES dataset, which is used for seizure prediction, the CHB-MIT dataset, primarily designed for seizure detection, requires additional data segmentation and annotation for prediction tasks. Moreover, as the CHB-MIT dataset contains sEEG recordings with a lower signal-to-noise ratio compared to iEEG recordings in the AES dataset, noise suppression is essential during preprocessing.

Then, Section 5.3.3 in Chapter 5 introduces a comprehensive on-chip energy model that assesses the energy costs during the inference phase, including the compute energy, weight access energy, and activation access energy.

The refined fixed-precision quantization scheme achieves better results than those detailed in Chapter 4 due to improvements in the scaling factor, quantization granularity and training strategy. In Chapter 4, the scaling factors for the weights, biases, and activations were set based on a per-layer style. However, in Chapter 5, while the scaling factors for biases and activations remain per layer, the scaling factors for weights are adjusted based on a per-channel style. This adjustment helps minimize rounding and clipping errors, leading to improved 16-bit and 8-bit quantization performance in Chapter 5 compared to Chapter 4, as illustrated in Table 8.1. The performance of 4-bit, 2-bit, and 1-bit quantization cannot be improved due to the significant error. Both Chapter 4 and Chapter 5 employ an identical training approach for the 32-bit floating-point baseline model. In Chapter 4, this training strategy is extended to the quantization model, using the Adam optimizer for 100 epochs with β_1 set to 0.9, β_2 set to 0.999, and an initial learning rate of 0.001 that is reduced by half every 50 epochs. In contrast, Chapter 5 also utilizes the Adam optimizer with the same β_1 and β_2 values for training but extends the process to 200 epochs, maintaining the same initial learning rate of 0.001, which is similarly halved every 50 epochs. This adjustment is due to the quantization model, which demands more time and fine-tuning of the learning rate to achieve optimal performance. Specifically in Table 8.1, the 1DSCNN baseline model in Chapter 5 achieves a sensitivity of 94.44%, a FPR of 0.011/h, an AUC of 0.979, a model size of 21.32 kB, and an estimated energy consumption of 0.20 μJ for the AES dataset.

In Chapter 6, inspired by the attention mechanism and 1DSCNN proposed in Chapter 4 and Chapter 5, an attention-driven $S^3$1DCNN is presented to precisely predict epileptic seizures, demonstrating its ability to interpret and analyze spatio-temporal, non-stationary iEEG recordings for precise detection of epilepsy onset areas. The main difference between this chapter and Chapter 4 and 5 is that the convolutional layer in Figure 5.3 is replaced by the spectral-spatial attention module. The computational flow of the spectral-spatial attention module is detailed in Section 7.2.2. This module integrates spectral and spatial information from its local receptive field, thereby creating improved feature representations by weighting different features' importance dynamically.

Meanwhile, this attention module can seamlessly incorporate with Grad-CAM to identify the channels and frequency components that most significantly influence prediction outcomes, which can help identify the location of seizure onset. This is the first epileptic seizure prediction method to combine an attention mechanism with Grad-CAM, allowing the identification of key input features that contribute to the model prediction. In addition, the on-chip energy model presented in Chapter 5 is also employed to access the energy costs of this model. Specifically in Table 8.1, compared to the 32-bit floating-point baseline models (1DSCNN) in Chapter 4 and Chapter 5, the proposed $S^3$1DCNN in Chapter 6 achieves a slightly better AUC of 0.980, while maintaining a 0.84% larger model size of 21.50 kB with a 5.1% higher energy consumption of 0.208 μJ energy consumption. Still, it offers interpretability to clarify the seizure onset zone. This feature can offer healthcare professionals valuable insights into seizure onset regions for better preventive measures and medication interventions.

Table 8.2 compares the results of the second research objective with the recent state-of-the-art methods for the CHB-MIT dataset. All methods in Table 8.2 employ CNNs for classification except [43] and [78], and the majority of methods [19, 34, 75, 77, 78] adopt a preictal duration of 30 minutes. Regarding the CHB-MIT dataset, it should be noticed that various methods adopt different numbers of patients. Consequently, the 10 patients with the highest AUC scores are chosen from the 24 patients for a more equitable performance comparison with [19, 43]. The proposed 32-bit floating-point baseline model with 10 patients achieves sensitivity of 97.92%, FPR of 0.003/h, and AUC of 1 comparable to [19, 43], while achieving a reduction of 3.03 and 1.91 times in model size compared to each, respectively. For full 24 patients, the proposed baseline model achieves the best sensitivity of 96.14%, FPR of 0.018/h, AUC of 0.996 with the smallest model size of 23.62 kB and estimated energy consumption of 0.28 μJ when compared to [34, 75, 77, 78]. Table 8.1 and Table 8.2 validate the generalization ability of the proposed baseline model (1DSCNN) and quantization models in Chapter 5 on the AES and CHB-MIT datasets.

This thesis compares performance and energy consumption of the proposed models only to architectures based on DL methods, such as CNNs, to ensure a fair and straightforward comparison of energy efficiency. The focus on DL-based methods aligns with the energy model's design, which was developed with DL considerations. However, it is important to note that energy-efficient methods for seizure prediction are not limited to DL. For example, Hsieh et al. [103] introduced the world's first energy-efficient and real-time neural signal processor for seizure prediction with a reconfigurable SVM, which achieved a 92.0%

sensitivity, a 0.57/h FPR and a 96.2 nJ/class energy consumption on the CHB-MIT dataset. The processor shows state-of-the-art performance in both seizure prediction and detection, which marks a significant advancement in implantable devices for epilepsy management. As their excellent results are based on measurements performed on a fabricated ASIC, it is hard to compare them fairly with those reported in the present thesis. It is also worth mentioning that their results were published in 2023, after some of our core contributions were published in 2022.

8.4 Implications and Significance

The development of energy-efficient and interpretable seizure prediction methods holds significant implications for both patients with epilepsy and the broader medical community. Energy efficiency in these systems is crucial, particularly for wearable devices that monitor signs of impending seizures. These devices must operate continuously; therefore, using minimal power extends battery life, enhancing user convenience and ensuring constant monitoring without frequent recharges.

Furthermore, interpretable prediction methods are essential for gaining trust and broader acceptance among healthcare providers and patients. When the mechanisms behind predictions are transparent, clinicians are better equipped to make informed decisions based on system alerts. Patients, on their part, can feel more confident in managing their condition with devices that provide understandable and reliable warning signs.

These advances could transform the management and treatment of epilepsy, making proactive care more accessible and reducing the unpredictability associated with seizures. Ultimately, this would improve patients' quality of life, as they could engage in daily activities with greater safety and fewer interruptions.

CHAPTER 9 CONCLUSION

Seizure prediction is crucial for individuals with drug-resistant epilepsy, who may not respond well to AEDs. For these patients, it significantly enhances safety by allowing individuals to take precautionary measures or alternative interventions against potential injuries associated with unexpected seizures. In addition, seizure prediction helps minimize the reliance on medications that can have severe side effects. Many AEDs come with a range of side effects, including cognitive impairment, mood disturbances, and physical health problems, which can affect patient quality of life. By enabling better timing and reducing medication dosage through precise prediction, patients can experience fewer side effects while maintaining effective seizure control. Thus, wearable or implantable devices are critical in epilepsy management because they provide continuous and real-time monitoring of EEG signals. This capability allows the detection of precursory seizure patterns, enabling early warnings and interventions that can improve patient autonomy and quality of life.

9.1 Summary of Works

This thesis focuses on a robust epileptic seizure prediction engine tailored for wearable medical devices from algorithm perspective to explore practical approaches to help epileptic patients obtain seizure prediction information in time. This work's core aim is to develop high-performance, energy-efficient, and interpretable methods for seizure prediction in resource-limited wearable devices that can potentially be suitable for daily use. The main challenge of this work lies in the trade-off between performance and energy consumption, highlighting an ongoing dilemma in resource-limited wearable medical devices. Although the ultimate goal is implementation on wearable platforms, this work focuses on algorithms and examines their energy efficiency as a step toward solutions suited for resource-limited systems. Furthermore, the core aim of this work can be divided into three objectives, as mentioned in Section 1.3. The thesis contributions are summarized again as follows:

- The first objective was to design a high-performance, patient-specific CNN with tiny model size for effective seizure prediction, which helps alleviate patient anxiety (low FPR) and enable preventive measures or medication interventions (high sensitivity). To achieve the first objective, a 1DSCNN is proposed to predict epilepsy seizures. This model is designed with a very competitive small size suitable for wearable biomedical devices. Compared to state-of-the-art methods, the proposed 1DSCNN achieves the best performance with the smallest model size on the AES dataset. In addition, a preliminary quantization scheme is also applied to evaluate the impact of various bit widths on model performance, facilitating its deployment in wearable biomedical devices.

- The second objective was to further explore an energy-efficient, patient-specific DL algorithm for effectively predicting epileptic seizures, addressing the trade-off between performance and energy consumption in resource-limited wearable medical devices. To achieve the second objective, firstly, our proposed 1DSCNN in the first objective is further evaluated to demonstrate its generalization on the AES and CHB-MIT datasets, which surpasses state-of-the-art methods in terms of sensitivity, FPR, AUC, model size and energy consumption. Then, a hardware-friendly, fixed-precision quantization scheme has been implemented on the two datasets, resulting in excellent energy efficiency with minimal performance losses. Finally, a near-optimal, low-complexity mixed-precision search method is also suggested, which has the potential to improve model performance for subjects with epileptic seizures that are challenging to predict.

- The third objective was to develop a high-performance, interpretable CNN with very low energy consumption for accurately forecasting epileptic seizures, which is intended to be transparent and understandable to healthcare professionals. To achieve the third objective, an attention-driven $S^3$1DCNN is proposed to forecast seizures, featuring its interpretable ability to analyze spatiotemporal non-stationary iEEG recordings for precisely localizing regions of epilepsy onset. In addition, the proposed $S^3$1DCNN outperforms state-of-the-art methods regarding AUC, model size, and energy consumption on the AES dataset, which shows its excellent potential for low-power biomedical wearable devices.

9.2 Limitations

9.2.1 EEG Recordings

In this thesis, both iEEG and sEEG datasets are employed to predict epileptic seizures. While these EEGs offer the advantage of high temporal resolution, capturing the rapid dynamics of brain activity during seizures, they are insufficient to provide adequate spatial resolution. This limitation is significant as it limits the ability to localize the origin of seizure activity within the brain precisely. Accurate localization is essential to both predict seizures and target treatments effectively in the clinical management of epilepsy.

9.2.2 Patient-Specific Algorithm

The patient-specific algorithm for epileptic seizure prediction has several critical limitations. The lack of generalizability in patient-specific models requires each to be developed and maintained individually. Furthermore, the approach is notably limited by the scarcity of preictal recordings and the imbalance between preictal and interictal states. Moreover, the approach demands prolonged and continuous individualized EEG signals. This is due to the variability in seizure patterns between individuals and within the same patient over time, which further requires developing and ongoing tuning prediction models. Thus, these factors underscore the significant challenges and limitations inherent in the patient-specific approach to predict epileptic seizures.

9.2.3 Insufficient Interpretability

Although Grad-CAM is useful for interpretability, it provides limited insights into the prediction process of a model. Grad-CAM highlights the regions of the input that are most active in a particular output class; it does not explain intermediate layers and causality. It cannot determine whether the highlighted regions directly caused the prediction or are correlated with it. Meanwhile, it does not explain whether the model focuses on specific frequency patterns, spatial correlations, or noise.

9.3 Future Research

9.3.1 High-Density EEG

High-density electrode grids consist of a significantly increased number of electrodes compared to standard setups, providing much finer spatial resolution to obtain a more detailed and precise brain activity mapping. For example, g.Pangolin [83] reported the world's first ultra-high density non-invasive electrode grids featuring 1024 channels for sEEG recording. This enhanced resolution could lead to more accurate and earlier seizure prediction, improved localization of seizure onset regions, and better epilepsy treatments.

9.3.2 Lifelong Learning

The limitations of the patient-specific algorithm can be addressed by incorporating lifelong learning into patient-specific wearable devices. Lifelong learning is designed to adapt continuously, learning from new data without forgetting previous knowledge. This approach is particularly suited to the dynamic nature of epilepsy, where the seizure patterns of a patient may evolve. Integrating lifelong learning into wearable devices could perpetually refine and update patient-specific models based on ongoing patient EEG signals.

9.3.3 Foundation Model

The scarcity of preictal recordings could benefit significantly from adopting a foundation model approach, where the model is trained on vast public EEG datasets to develop a broad understanding before being fine-tuned for specific patients. The foundation model can capture a wide range of seizure patterns to enhance generalizability and robustness compared to patient-specific models. Furthermore, once a foundation model is developed, it can be quickly adapted to new patients using a smaller subset of their EEG data, significantly reducing the time and resources required for model deployment. This approach reduces the need for long-term monitoring and enhances the scalability of seizure prediction, potentially making robust predictions accessible to a broader range of epilepsy patients.

9.3.4 Data Security and Privacy

As prediction models rely on continuous and real-time neurological EEG signals collection from wearable devices or implantable sensors, ensuring the integrity and confidentiality of these EEG signals becomes essential. Advanced encryption methods, secure transmission protocols, and robust access controls should be integral to research and development efforts. These measures will protect patient information and encourage wider adoption and participation in long-term monitoring programs essential for improving seizure prediction models.

9.4 Knowledge Dissemination

The methodologies and results developed in this thesis have been disseminated to the academic community through two conference oral presentations, one conference poster presentation, and one journal paper submission, as detailed below:

- **Y. Zhang**, Y. Savaria, M. Sawan, and F. Leduc-Primeau, "Tiny Neural Network for Epileptic Seizure Forecasting in Wearable Devices," *IEEE Transactions on Biomedical Engineering*, submitted, 2024/03/20.
- **Y. Zhang**, Y. Savaria, M. Sawan, and F. Leduc-Primeau, "An energy-efficient neural network for seizure prediction and localization with wearables," in *7th edition of the Montreal AI and Neuroscience conference*, 22-25 Oct. 2024, Montreal. (Poster presentation)

- **Y. Zhang**, Y. Savaria, M. Sawan, and F. Leduc-Primeau, "$S^3$1DCNN: A compact stacked spectral-spatial attention 1DCNN for seizure prediction with wearables," in *2024 22st IEEE Interregional NEWCAS Conference (NEWCAS).* IEEE, 2024, pp. 278-282, doi: 10.1109/NewCAS58973.2024.10666297. (Oral presentation)

- **Y. Zhang**, Y. Savaria, S. Zhao, G. Mordido, M. Sawan, and F. Leduc-Primeau, "Tiny cnn for seizure prediction in wearable biomedical devices," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).* IEEE, 2022, pp. 1306-1309, doi: 10.1109/EMBC48229.2022.9872006. (Oral presentation)

- **Y. Zhang**, Y. Savaria, S. Zhao, G. Mordido, M. Sawan, and F. Leduc-Primeau, "Quantized one-dimensional stacked cnn for seizure forecasting with wearables," in *2022 Edge Intelligence Workshop*, 19-20 Sep. 2022, Montreal. (Poster presentation)

REFERENCES

[1] R. Shriram, M. Sundhararajan, and N. Daimiwal, "Eeg based cognitive workload assessment for maximum efficiency," *Int. Organ. Sci. Res. IOSR*, vol. 7, pp. 34–38, 2013.

[2] J. Parvizi and S. Kastner, "Promises and limitations of human intracranial electroencephalography," *Nature neuroscience*, vol. 21, no. 4, pp. 474–483, 2018.

[3] P. Kwan, A. Arzimanoglou, A. T. Berg, M. J. Brodie, W. Allen Hauser, G. Mathern, S. L. Moshé, E. Perucca, S. Wiebe, and J. French, "Definition of drug resistant epilepsy: consensus proposal by the ad hoc task force of the ilae commission on therapeutic strategies," 2010.

[4] L. Kuhlmann, K. Lehnertz, M. P. Richardson, B. Schelter, and H. P. Zaveri, "Seizure prediction—ready for a new era," *Nature Reviews Neurology*, vol. 14, no. 10, pp. 618–630, 2018.

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

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

[7] World Health Organization, *Epilepsy: a public health imperative.* World Health Organization, 2019.

[8] J. Engel, *Seizures and epilepsy.* Oxford University Press, USA, 2013, vol. 83.

[9] G. A. Baker, J. Brooks, D. Buck, and A. Jacoby, "The stigma of epilepsy: a european perspective," *Epilepsia*, vol. 41, no. 1, pp. 98–104, 2000.

[10] G. A. Baker, A. Jacoby, D. Buck, C. Stalgis, and D. Monnet, "Quality of life of people with epilepsy: a european study," *Epilepsia*, vol. 38, no. 3, pp. 353–362, 1997.

[11] M. Leitinger, E. Trinka, E. Gardella, A. Rohracher, G. Kalss, E. Qerama, J. Höfler, A. Hess, G. Zimmermann, G. Kuchukhidze *et al.*, "Diagnostic accuracy of the salzburg eeg criteria for non-convulsive status epilepticus: a retrospective study," *The Lancet Neurology*, vol. 15, no. 10, pp. 1054–1062, 2016.

[12] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields.* Lippincott Williams & Wilkins, 2005.

[13] W. O. Tatum IV, *Handbook of EEG interpretation.* Springer Publishing Company, 2021.

[14] B. C. Jobst, F. Bartolomei, B. Diehl, B. Frauscher, P. Kahane, L. Minotti, A. Sharan, N. Tardy, G. Worrell, and J. Gotman, "Intracranial eeg in the 21st century," *Epilepsy currents*, vol. 20, no. 4, pp. 180–188, 2020.

[15] D. Whitmer, G. Worrell, M. Stead, I. K. Lee, and S. Makeig, "Utility of independent component analysis for interpretation of intracranial eeg," *Frontiers in human neuroscience*, vol. 4, p. 184, 2010.

[16] M. Bandarabadi, "Low-complexity measures for epileptic seizure prediction and early detection based on classification," Ph.D. dissertation, Universidade de Coimbra (Portugal), 2015.

[17] M. J. Cook, T. J. O'Brien, S. F. Berkovic, M. Murphy, A. Morokoff, G. Fabinyi, W. D'Souza, R. Yerra, J. Archer, L. Litewka *et al.*, "Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study," *The Lancet Neurology*, vol. 12, no. 6, pp. 563–571, 2013.

[18] S. Beniczky, P. Karoly, E. Nurse, P. Ryvlin, and M. Cook, "Machine learning and wearable devices of the future," *Epilepsia*, vol. 62, pp. S116–S124, 2021.

[19] S. Zhao, J. Yang, and M. Sawan, "Energy-efficient neural network for epileptic seizure prediction," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 1, pp. 401–411, 2021.

[20] D. R. Freestone, P. J. Karoly, and M. J. Cook, "A forward-looking review of seizure prediction," *Current opinion in neurology*, vol. 30, no. 2, pp. 167–173, 2017.

[21] S. S. Viglione and G. O. Walsh, "Proceedings: Epileptic seizure prediction," *Electroencephalography clin. neurophysiol.*, vol. 39, no. 4, pp. 435–436, 1975.

[22] Z. Rogowski, I. Gath, and E. Bental, "On the prediction of epileptic seizures," *Biological cybernetics*, vol. 42, no. 1, pp. 9–15, 1981.

[23] J. Martinerie, C. Adam, M. Quyen, M. Baulac, S. Clemenceau, B. Renault, and F. J. Varela, “Epileptic seizures can be anticipated by non-linear analysis,” *Nature medicine*, vol. 4, no. 10, pp. 1173–1176, 1998.

[24] K. Lehnertz and B. Litt, “The first international collaborative workshop on seizure prediction: summary and data description,” *Clinical neurophysiology*, vol. 116, no. 3, pp. 493–505, 2005.

[25] W. van Drongelen, S. Nayak, D. M. Frim, M. H. Kohrman, V. L. Towle, H. C. Lee, A. B. McGee, M. S. Chico, and K. E. Hecox, “Seizure anticipation in pediatric epilepsy: use of kolmogorov entropy,” *Pediatric neurology*, vol. 29, no. 3, pp. 207–213, 2003.

[26] F. Mormann, T. Kreuz, R. G. Andrzejak, P. David, K. Lehnertz, and C. E. Elger, “Epileptic seizures are preceded by a decrease in synchronization,” *Epilepsy research*, vol. 53, no. 3, pp. 173–185, 2003.

[27] M. Ihle, H. Feldwisch-Drentrup, C. A. Teixeira, A. Witon, B. Schelter, J. Timmer, and A. Schulze-Bonhage, “Epilepsiae–a european epilepsy database,” *Computer methods and programs in biomedicine*, vol. 106, no. 3, pp. 127–138, 2012.

[28] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, “Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state,” *Physical Review E*, vol. 64, no. 6, p. 061907, 2001.

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

[30] P. Thodoroff, J. Pineau, and A. Lim, “Learning robust features using deep learning for automatic seizure detection,” in *Machine learning for healthcare conference*. PMLR, 2016, pp. 178–190.

[31] O. Ouichka, A. Echtioui, and H. Hamam, “Deep learning models for predicting epileptic seizures using ieeg signals,” *Electronics*, vol. 11, no. 4, p. 605, 2022.

[32] L. Kuhlmann, P. Karoly, D. R. Freestone, B. H. Brinkmann, A. Temko, A. Barachant, F. Li, G. Titericz Jr, B. W. Lang, D. Lavery *et al.*, “Epilepsyecosystem. org: crowd-sourcing reproducible seizure prediction with long-term human intracranial eeg,” *Brain*, vol. 141, no. 9, pp. 2619–2630, 2018.

[33] K. M. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals," *Computers in biology and medicine*, vol. 99, pp. 24–37, 2018.

[34] N. D. Truong, A. D. Nguyen, L. Kuhlmann, M. R. Bonyadi, J. Yang, S. Ippolito, and O. Kavehei, "Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram," *Neural Networks*, vol. 105, pp. 104–111, 2018.

[35] L. A. S. Kitano, M. A. A. Sousa, S. D. Santos, R. Pires, S. Thome-Souza, and A. B. Campo, "Epileptic seizure prediction from eeg signals using unsupervised learning and a polling-based decision process," in *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part II 27.* Springer, 2018, pp. 117–126.

[36] O. Stojanović, L. Kuhlmann, and G. Pipa, "Predicting epileptic seizures using nonnegative matrix factorization," *PloS one*, vol. 15, no. 2, p. e0228025, 2020.

[37] R. Hussein, S. Lee, R. Ward, and M. J. McKeown, "Semi-dilated convolutional neural networks for epileptic seizure prediction," *Neural Networks*, vol. 139, pp. 212–222, 2021.

[38] I. Korshunova, P.-J. Kindermans, J. Degrave, T. Verhoeven, B. H. Brinkmann, and J. Dambre, "Towards improved design and evaluation of epileptic seizure predictors," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 3, pp. 502–510, 2017.

[39] N. D. Truong, L. Kuhlmann, M. R. Bonyadi, D. Querlioz, L. Zhou, and O. Kavehei, "Epileptic seizure forecasting with generative adversarial networks," *IEEE Access*, vol. 7, pp. 143 999–144 009, 2019.

[40] Y. Yang, M. Zhou, Y. Niu, C. Li, R. Cao, B. Wang, P. Yan, Y. Ma, and J. Xiang, "Epileptic seizure prediction based on permutation entropy," *Frontiers in computational neuroscience*, vol. 12, p. 55, 2018.

[41] S. M. Usman, S. Khalid, and M. H. Aslam, "Epileptic seizures prediction using deep learning techniques," *Ieee Access*, vol. 8, pp. 39 998–40 007, 2020.

[42] C. Li, C. Shao, R. Song, G. Xu, X. Liu, R. Qian, and X. Chen, "Spatio-temporal mlp network for seizure prediction using eeg signals," *Measurement*, vol. 206, p. 112278, 2023.

[43] H. Daoud and M. A. Bayoumi, "Efficient epileptic seizure prediction based on deep learning," *IEEE transactions on biomedical circuits and systems*, vol. 13, no. 5, pp. 804–813, 2019.

[44] Y. Xu, J. Yang, S. Zhao, H. Wu, and M. Sawan, "An end-to-end deep learning approach for epileptic seizure prediction," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS).* IEEE, 2020, pp. 266–270.

[45] X. Wu, Z. Yang, T. Zhang, L. Zhang, and L. Qiao, "An end-to-end seizure prediction approach using long short-term memory network," *Frontiers in Human Neuroscience*, vol. 17, p. 1187794, 2023.

[46] S. M. Usman, S. Latif, and A. Beg, "Principle components analysis for seizures prediction using wavelet transform," *arXiv preprint arXiv:2004.07937*, 2020.

[47] H.-H. Chen, H.-T. Shiao, and V. Cherkassky, "Online prediction of lead seizures from ieeg data," *Brain Sciences*, vol. 11, no. 12, p. 1554, 2021.

[48] J. S. Ra, T. Li, and Y. Li, "A novel permutation entropy-based eeg channel selection for improving epileptic seizure prediction," *Sensors*, vol. 21, no. 23, p. 7972, 2021.

[49] M. M. Qureshi and M. Kaleem, "Eeg-based seizure prediction with machine learning," *Signal, Image and Video Processing*, vol. 17, no. 4, pp. 1543–1554, 2023.

[50] G. Costa, C. Teixeira, and M. F. Pinto, "Comparison between epileptic seizure prediction and forecasting based on machine learning," *Scientific Reports*, vol. 14, no. 1, p. 5653, 2024.

[51] D. Lee, B. Kim, T. Kim, I. Joe, J. Chong, K. Min, and K. Jung, "A resnet-lstm hybrid model for predicting epileptic seizures using a pretrained model with supervised contrastive learning," *Scientific Reports*, vol. 14, no. 1, p. 1319, 2024.

[52] S. Shi and W. Liu, "B2-vit net: Broad vision transformer network with broad attention for seizure prediction," *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 32, pp. 178–188, 2024.

[53] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, *Efficient processing of deep neural networks.* Springer, 2020.

[54] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort, "A white paper on neural network quantization," *arXiv preprint arXiv:2106.08295*, 2021.

[55] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[58] M. Black and D. Graham, "Sudden death in epilepsy," *Current Diagnostic Pathology*, vol. 8, no. 6, pp. 365–372, 2002.

[59] Y. Yuan, G. Xun, K. Jia, and A. Zhang, "A multi-view deep learning framework for eeg seizure detection," *IEEE journal of biomedical and health informatics*, vol. 23, no. 1, pp. 83–94, 2018.

[60] A. Aarabi and B. He, "A rule-based seizure prediction method for focal neocortical epilepsy," *Clinical Neurophysiology*, vol. 123, no. 6, pp. 1111–1122, 2012.

[61] C.-L. Liu, B. Xiao, W.-H. Hsaio, and V. S. Tseng, "Epileptic seizure prediction with multi-view convolutional neural networks," *IEEE access*, vol. 7, pp. 170 352–170 361, 2019.

[62] American epilepsy society seizure prediction challenge. [Online]. Available: www.kaggle.com/c/seizure-prediction/data

[63] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction.* Springer, 2009, vol. 2.

[64] N. Czarnek, K. Morton, L. Collins, S. Tantum, and C. Throckmorton, "The impact of time on seizure prediction performance in the fspeeg database," *Epilepsy & Behavior*, vol. 48, pp. 79–82, 2015.

[65] B. H. Brinkmann, J. Wagenaar, D. Abbot, P. Adkins, S. C. Bosshard, M. Chen, Q. M. Tieng, J. He, F. Muñoz-Almaraz, P. Botella-Rocamora *et al.*, "Crowdsourcing reproducible seizure forecasting in human and canine epilepsy," *Brain*, vol. 139, no. 6, pp. 1713–1722, 2016.

[66] A. V. Oppenheim, J. R. Buck, and R. W. Schafer, *Discrete-time signal processing. Vol. 2.* Upper Saddle River, NJ: Prentice Hall, 2001.

[67] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[68] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[69] X. Lin, C. Zhao, and W. Pan, "Towards accurate binary convolutional neural network," *Advances in neural information processing systems*, vol. 30, 2017.

[70] C. N. Coelho, A. Kuusela, S. Li, H. Zhuang, J. Ngadiuba, T. K. Aarrestad, V. Loncar, M. Pierini, A. A. Pol, and S. Summers, "Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors," *Nature Machine Intelligence*, vol. 3, no. 8, pp. 675–686, 2021.

[71] B. Moons, K. Goetschalckx, N. Van Berckelaer, and M. Verhelst, "Minimum energy quantized neural networks," in *2017 51st Asilomar Conference on Signals, Systems, and Computers.* IEEE, 2017, pp. 1921–1925.

[72] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6869–6898, 2017.

[73] 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*, 2017, pp. 213–222.

[74] M. Ma, Y. Cheng, Y. Wang, X. Li, Q. Mao, Z. Zhang, Z. Chen, and Y. Zhou, "Early prediction of epileptic seizure based on the bnlstm-casa model," *IEEE access*, vol. 9, pp. 79 600–79 610, 2021.

[75] D. Wu, Y. Shi, Z. Wang, J. Yang, and M. Sawan, "C^{2}SP-Net: joint compression and classification network for epilepsy seizure prediction," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 841–850, 2023.

[76] J. Kern, S. Henwood, G. Mordido, E. Dupraz, A. Aïssa-El-Bey, Y. Savaria, and F. Leduc-Primeau, "Fast and accurate output error estimation for memristor-based deep neural networks," *IEEE Transactions on Signal Processing*, 2024.

[77] Y. Zhang, Y. Guo, P. Yang, W. Chen, and B. Lo, "Epilepsy seizure prediction on eeg using common spatial pattern and convolutional neural network," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 465–474, 2019.

[78] L. Tang, N. Xie, M. Zhao, and X. Wu, "Seizure prediction using multi-view features and improved convolutional gated recurrent network," *IEEE Access*, vol. 8, pp. 172 352–172 361, 2020.

[79] D. Liang, A. Liu, Y. Gao, C. Li, R. Qian, and X. Chen, "Semi-supervised domain-adaptive seizure prediction via feature alignment and consistency regularization," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.

[80] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[81] S. Kovac, V. N. Vakharia, C. Scott, and B. Diehl, "Invasive epilepsy surgery evaluation," *Seizure*, vol. 44, pp. 125–136, 2017.

[82] A. K. Shah and S. Mittal, "Invasive electroencephalography monitoring: Indications and presurgical planning," *Annals of Indian Academy of Neurology*, vol. 17, no. Suppl 1, p. S89, 2014.

[83] "g.PANGOLIN Ultra High-Density EEG/EMG/ECG | g.tec medical engineering GmbH — gtec.at," https://www.gtec.at/product/g-pangolin-electrodes/, [Accessed 28-Sep-2022].

[84] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "14.5 envision: A 0.26-to-10tops/w subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm fdsoi," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2017, pp. 246–247.

[85] H. Khan, L. Marcuse, M. Fields, K. Swann, and B. Yener, “Focal onset seizure prediction using convolutional networks,” *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 2109–2118, 2017.

[86] K. Fei, W. Wang, Q. Yang, and S. Tang, “Chaos feature study in fractional fourier domain for preictal prediction of epileptic seizure,” *Neurocomputing*, vol. 249, pp. 290–298, 2017.

[87] Y. Zhang, Y. Savaria, S. Zhao, G. Mordido, M. Sawan, and F. Leduc-Primeau, “Tiny cnn for seizure prediction in wearable biomedical devices,” in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 1306–1309.

[88] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, “1d convolutional neural networks and applications: A survey,” *Mechanical systems and signal processing*, vol. 151, p. 107398, 2021.

[89] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[90] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, “A survey of quantization methods for efficient neural network inference,” *arXiv preprint arXiv:2103.13630*, 2021.

[91] Y. Bengio, N. Léonard, and A. Courville, “Estimating or propagating gradients through stochastic neurons for conditional computation,” *arXiv preprint arXiv:1308.3432*, 2013.

[92] K. Wang, Z. Liu, Y. Lin, J. Lin, and S. Han, “Haq: Hardware-aware automated quantization with mixed precision,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8612–8620.

[93] B. Wu, Y. Wang, P. Zhang, Y. Tian, P. Vajda, and K. Keutzer, “Mixed precision quantization of convnets via differentiable neural architecture search,” *arXiv preprint arXiv:1812.00090*, 2018.

[94] H. Yang, L. Duan, Y. Chen, and H. Li, “Bsq: Exploring bit-level sparsity for mixed-precision neural network quantization,” *arXiv preprint arXiv:2102.10462*, 2021.

[95] A. Ng, “Machine learning yearning,” *URL: http://www. mlyearning. org/(96)*, vol. 139, 2017.

[96] M. AskariHemmat, R. A. Hemmat, A. Hoffman, I. Lazarevich, E. Saboori, O. Mastropietro, S. Sah, Y. Savaria, and J.-P. David, “Qreg: On regularization effects of quantization,” *arXiv preprint arXiv:2206.12372*, 2022.

[97] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh *et al.*, “Mixed precision training,” *arXiv preprint arXiv:1710.03740*, 2017.

[98] Ö. Yıldırım, U. B. Baloglu, and U. R. Acharya, “A deep convolutional neural network model for automated identification of abnormal eeg signals,” *Neural Computing and Applications*, vol. 32, pp. 15 857–15 868, 2020.

[99] X. Wang, X. Wang, W. Liu, Z. Chang, T. Kärkkäinen, and F. Cong, “One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial eeg,” *Neurocomputing*, vol. 459, pp. 212–222, 2021.

[100] Z. Wang, J. Yang, H. Wu, J. Zhu, and M. Sawan, “Power efficient refined seizure prediction algorithm based on an enhanced benchmarking,” *Scientific Reports*, vol. 11, no. 1, p. 23498, 2021.

[101] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[102] Y. Sato, S. M. Wong, Y. Iimura, A. Ochi, S. M. Doesburg, and H. Otsubo, “Spatiotemporal changes in regularity of gamma oscillations contribute to focal ictogenesis,” *Scientific reports*, vol. 7, no. 1, p. 9362, 2017.

[103] Y.-Y. Hsieh, Y.-C. Lin, and C.-H. Yang, “A 96.2-nj/class neural signal processor with adaptable intelligence for seizure prediction,” *IEEE Journal of Solid-State Circuits*, vol. 58, no. 1, pp. 167–176, 2022.