Degree Project in Technology

Second cycle, 30 credits

Computationally Efficient and Generalizable Machine Learning Algorithms for Seizure Detection from EEG Signals

ZHEYUN SHOU

Computationally Efficient and Generalizable Machine Learning Algorithms for Seizure Detection from EEG Signals

ZHEYUN SHOU

Master's Programme, Systems, Control and Robotics, 120 credits
Date: July 1, 2025

Supervisors: Erik Fransén, Gonzalo Uribarri
Examiner: Arvind Kumar
School of Electrical Engineering and Computer Science
Swedish title: Beräkningseffektiva och Generaliserbara Maskininlärningsalgoritmer för Anfallsdetektering från EEG-signaler

Abstract

Automated seizure detection from Electroencephalography (EEG) is critical for managing epilepsy, a neurological disorder affecting millions worldwide. Manual EEG interpretation is time-consuming and subjective, creating a need for algorithms that are not only accurate but also computationally efficient and generalizable across diverse patient populations. To address this challenge, this thesis investigates the Detach-ROCKET framework by implementing and evaluating the established Detach Ensemble method alongside a single Detach-MINIROCKET model. These models are benchmarked against the catch22 classifier on the public TUSZ and Siena EEG datasets, with the event-wise analysis adhering to the scoring standards of the 2025 Seizure Detection Challenge. Results show that the Detach Ensemble provides superior performance, achieving a high event-wise F1-score (0.89 on the TUSZ test set, 0.39 on cross-dataset evaluation) and proving remarkable inference efficiency. These findings establish the Detach Ensemble as a powerful and practical framework for seizure detection, offering an effective balance of predictive accuracy, speed, and cross-dataset generalizability suitable for clinical applications.

Keywords

Seizure Detection, Electroencephalography (EEG), Machine Learning, Time Series Classification, Detach-ROCKET

Sammanfattning

Automatiserad anfallsdetektering från elektroencefalografi (EEG) är avgörande för behandling av epilepsi, en neurologisk sjukdom som drabbar miljontals människor världen över. Manuell EEG-tolkning är tidskrävande och subjektiv, vilket skapar ett behov av algoritmer som inte bara är noggranna utan också beräkningsmässigt effektiva och generaliserbara över olika patientpopulationer. För att möta denna utmaning undersöker denna avhandling Detach-ROCKET-ramverket genom att implementera och utvärdera den etablerade Detach Ensemble-metoden tillsammans med en enstaka Detach-MINIROCKET-modell. Dessa modeller jämförs mot catch22-klassificeraren på de offentliga TUSZ- och Siena EEG-datamängderna, där den händelsebaserade analysen följer bedömningsstandarderna för 2025 Seizure Detection Challenge. Resultaten visar att Detach Ensemble ger överlägsen prestanda och uppnår höga händelsebaserade F1-poäng (0,89 på TUSZ-testuppsättningen, 0,39 på tvärdata-utvärdering) samt påvisar anmärkningsvärd inferenseffektivitet. Dessa fynd etablerar Detach Ensemble som ett kraftfullt och praktiskt ramverk för anfallsdetektering, vilket erbjuder en effektiv balans mellan prediktiv noggrannhet, hastighet och generalisering över datamängder som är lämplig för kliniska tillämpningar.

Nyckelord

Anfallsdetektering, Elektroencefalografi (EEG), Maskininlärning, Tidsserieklassificering, Detach-ROCKET

Acknowledgments

I would like to thank my supervisors, Erik Fransen and Gonzalo Uribarri, for their invaluable guidance, support, and expert advice throughout this project. I also wish to thank my examiner, Arvind Kumar, for his constructive feedback and insightful suggestions that helped improve this thesis.

Stockholm, July 2025
Zheyun Shou

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Epilepsy is a chronic neurological disorder characterized by an enduring predisposition to generate recurrent, unprovoked epileptic seizures—transient episodes of abnormal, excessive, or synchronous neuronal activity in the brain [1, 2]. This condition represents a significant global health challenge, affecting approximately 50 million individuals worldwide and ranking among the most prevalent serious neurological disorders [3]. Seizures can manifest through diverse symptoms ranging from subtle behavioral changes and brief lapses in awareness to dramatic convulsive episodes, depending on the brain regions involved and the pattern of electrical discharge propagation [4]. Beyond the immediate physical risks associated with seizure events, epilepsy imposes substantial burdens on patients through cognitive impairments, psychological comorbidities, and profound social limitations that collectively diminish quality of life [4, 5, 6]. The unpredictable nature of seizures necessitates continuous monitoring and rapid clinical response, making accurate and timely seizure detection a valuable technique for effective patient care, optimal therapeutic management, and the prevention of potentially life-threatening complications [7].

Electroencephalography (EEG) is the most widely used and reliable method for monitoring brain electrical activity and detecting seizures in clinical practice [8]. However, the manual interpretation of EEG recordings is a time-intensive process that requires specialized expertise, making it susceptible to inter-observer variability and potential delays in critical decision-making [9]. The increasing availability of long-term EEG monitoring systems has further amplified the volume of data requiring analysis, creating an urgent need for automated seizure detection algorithms that can assist clinicians in providing timely and accurate diagnoses.

Recent advances in machine learning and time series analysis have opened new possibilities for automated EEG analysis. Traditional approaches have relied heavily on handcrafted features and conventional machine learning algorithms [10, 11], while more recent developments have explored deep learning methodologies [12, 13]. However, many existing approaches face significant challenges in terms of computational efficiency, generalizability across diverse patient populations, and practical deployment in resource-constrained clinical environments.

The RandOm Convolutional KErnel Transform (ROCKET) family of algorithms has emerged as a promising paradigm for time series classification, offering state-of-the-art accuracy while maintaining exceptional computational efficiency [14, 15]. These methods transform time series data using randomly generated convolutional kernels, creating discriminative features that can be effectively utilized by simple linear classifiers. Recent extensions, including Detach-ROCKET [16], have further enhanced this approach through feature selection and ensemble methodologies.

1.1 Purpose

The primary purpose of this thesis is to address the persistent challenges of computational efficiency and model generalizability in the field of automated seizure detection from EEG signals. While existing methods have demonstrated varying degrees of success, many are either too computationally intensive for practical deployment or fail to perform reliably across different patient populations and datasets.

To this end, this study investigates the Detach-ROCKET algorithm and its variant, Detach Ensemble, to develop a seizure detection methodology. The goal is to achieve a compelling balance between high predictive performance, exceptional computational speed, and robust generalization, demonstrating a viable pathway toward a system practical for real-world clinical implementation, particularly in real-time monitoring and resource-constrained environments.

1.2 Research Question

This research addresses the following research questions:

To what extent can the Detach-ROCKET framework, particularly its ensemble variant, achieve a superior balance of predictive performance,

computational efficiency, and generalization for EEG seizure detection compared to established time-series classification benchmarks?

1.3 Objectives

To address the research purpose and answer the guiding research questions, the following are the main objectives of this thesis:

- To implement and adapt Detach-ROCKET and Detach Ensemble for the specific task of seizure detection from multivariate EEG signals.
- To establish a performance baseline by implementing and evaluating the `catch22` feature-based benchmark on the same datasets.
- To comprehensively evaluate and compare the Detach-ROCKET models against the benchmark using both epoch-wise and clinically-relevant event-wise metrics.
- To assess the generalization capabilities of the trained models by performing a cross-dataset evaluation on the unseen Siena Scalp EEG Database.
- To conduct a sensitivity analysis to quantify the impact of key hyperparameters on model performance.
- To investigate the interpretability of the ensemble model by analyzing the relative relevance of each EEG channel.
- To quantitatively evaluate the computational efficiency of each model by measuring and comparing their inference times on a representative EEG recording.

1.4 Delimitations

To maintain a clear and achievable scope, this study operates within several defined boundaries. The following points delineate the key delimitations of this research:

- Algorithmic Focus: This study is specifically focused on the application and evaluation of the Detach-ROCKET framework and its ensemble variant. While other methodologies like deep learning models (e.g.,

CNNs, Transformers) and traditional feature-based approaches are discussed in the background, they are not implemented or exhaustively compared against in the experimental sections. The evaluation is primarily benchmarked against `catch22` to provide a reference point as a state-of-the-art, feature-based time-series classification method.

- Dataset Scope: The research is conducted exclusively on two publicly available scalp EEG datasets: the Temple University Hospital EEG Seizure Corpus (TUSZ) and the Siena Scalp EEG Database. The findings and conclusions are therefore bound to the characteristics of this type of data and may not directly generalize to other forms of neurophysiological recordings, such as intracranial EEG (iEEG), neonatal EEG, or data from proprietary clinical systems with different acquisition protocols.
- Clinical Application Scope: The evaluation of the models is performed on pre-recorded and annotated data. This research does not extend to real-time implementation in a live clinical setting or the assessment of the algorithm's direct impact on patient management or clinical decision-making. The focus is on the technical validation of the algorithm's performance.
- Hyperparameter Optimization: Due to the computational and memory constraints discussed in the limitations, this study does not perform an exhaustive, grid-search-based optimization of all model hyperparameters. The sensitivity analysis is limited to a key subset of parameters (ensemble size, epoch duration, data size) to understand their general impact rather than to find a globally optimal configuration.

1.5 Structure of the thesis

The structure of this thesis is as follows: Chapter 2 provides the theoretical background on epilepsy, EEG, and relevant classification methods. Chapter 3 describes the methodology, including data processing, model implementation, and evaluation. Chapter 4 presents the results from the experiments. Chapter 5 discusses the findings and limitations of the study. Finally, Chapter 6 concludes the thesis with a summary of the contributions and directions for future work.

Chapter 2

Background

2.1 Epilepsy

A seizure represents a fundamental disruption of normal brain function, characterized by a transient surge of abnormal, excessive, or synchronous neuronal activity within the brain [1, 2]. These paroxysmal events can manifest through a wide spectrum of signs and symptoms, the nature of which is intrinsically linked to the specific cerebral networks involved in the ictal discharge and its subsequent propagation [4]. While an isolated seizure may be provoked by acute systemic insults or transient neurological stressors, the term epilepsy designates a more profound neurological condition: an enduring predisposition to generate unprovoked epileptic seizures, along with the associated neurobiological, cognitive, psychological, and social ramifications [17]. Clinically, epilepsy is typically diagnosed after at least one unprovoked seizure, especially when factors suggest a high likelihood of recurrence.

The global impact of seizures and epilepsy is considerable. Affecting an estimated 50 million individuals worldwide, epilepsy stands as one of the most prevalent serious neurological disorders [3]. Beyond the immediate risk of physical injury during an ictal event, seizures can precipitate or exacerbate cognitive difficulties, particularly in domains of memory and attention, and are frequently associated with psychological comorbidities such as anxiety and depression [5, 6]. The societal impact is also profound, with individuals often confronting stigma, discrimination, and limitations in education, employment, and personal autonomy [18]. The consequences of recurrent seizures permeate virtually every aspect of an individual's life.

The diagnostic process for seizures heavily relies on meticulous clinical

history-taking, including detailed eyewitness accounts, as the ictal events themselves are often brief and unpredictable, making direct observation by clinicians rare [19]. This dependence on subjective reporting, while indispensable, introduces potential inaccuracies and challenges in distinguishing seizures from other paroxysmal events [20]. Accurate seizure detection is not merely an academic exercise; it is critical for guiding appropriate therapeutic interventions, predicting long-term outcomes, and informing genetic counseling [7].

2.2 Electroencephalography for Brain Activity Monitoring

Electroencephalography (EEG) is a cornerstone neurophysiological technique used to record the electrical activity generated by the brain. It provides a non-invasive, direct measure of brain function by detecting voltage fluctuations resulting from ionic current flows within the neurons of the cerebral cortex [8, 21]. Specifically, EEG signals primarily reflect the summed postsynaptic potentials (both excitatory and inhibitory) of large populations of synchronously active pyramidal neurons oriented radially to the scalp [21, 22]. Due to its excellent temporal resolution, typically in the millisecond range, EEG is uniquely suited for capturing the rapidly changing dynamics of brain activity, making it an invaluable tool in both clinical neurology and neuroscience research, particularly for the assessment of conditions characterized by abnormal electrical discharges, such as epilepsy [23].

The acquisition of EEG data involves placing electrodes on the scalp, typically made of conductive materials like silver/silver-chloride (Ag/AgCl). These electrodes detect the minute electrical potentials (on the order of microvolts, μV) generated by the brain, which are then significantly amplified by differential amplifiers to make them suitable for digitization and subsequent analysis [24]. To ensure reproducibility and comparability of EEG recordings across different laboratories and individuals, standardized electrode placement systems are employed. The most widely adopted is the International 10-20 system and its extensions [25, 26]. This system positions electrodes at locations that are 10% or 20% of the total front-to-back or right-to-left distance of the skull, ensuring proportional spacing relative to cranial landmarks (nasion, inion, and preauricular points). Each electrode is labeled with a letter indicating the underlying brain lobe (e.g., F for frontal, P for parietal, T for temporal, O for occipital) and a number or another letter to denote its specific

position [25].

The resulting EEG waveforms are complex and are traditionally analyzed in terms of their frequency, amplitude, morphology, and topography. Clinically relevant information is often contained within specific frequency bands, such as delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (>30 Hz), each associated with different brain states or cognitive processes [21]. In the context of epilepsy, EEG is crucial for identifying interictal epileptiform discharges (IEDs) and for characterizing the electrographic signature of seizures themselves (ictal patterns) [27]. These patterns provide critical information for diagnosing epilepsy, classifying seizure types, localizing the seizure onset zone, and guiding treatment decisions.

However, the utility of EEG is often challenged by the inherent low signal-to-noise ratio (SNR) and its susceptibility to various artifacts [24, 28]. EEG signals are frequently contaminated by physiological artifacts originating from non-cerebral sources, such as eye movements and blinks (electrooculogram, EOG), muscle activity (electromyogram, EMG) particularly from scalp and facial muscles, and cardiac activity (electrocardiogram, ECG) [28, 29]. Non-physiological artifacts also pose significant problems, including 50/60 Hz power line interference, electrode impedance issues (e.g., electrode pop or poor contact), and movement artifacts from patient motion or cable sway [29]. These artifacts can mimic or obscure true epileptiform activity, leading to potential misinterpretations if not properly addressed. Consequently, rigorous data preprocessing, encompassing techniques such as filtering, artifact detection, and artifact removal or suppression, is an indispensable yet often complex step before any meaningful analysis or automated interpretation of EEG data can be performed [29, 30]. The challenge of reliably distinguishing pathological neural signals from this pervasive noise, especially in long-term recordings or in ambulatory settings, underscores the need for robust and sophisticated signal processing and machine learning algorithms, forming a critical motivation for the research presented in this thesis.

2.3 Time-series classification

Time Series Classification (TSC) is a specialized area within machine learning concerned with assigning predefined categorical labels to unlabeled time series data [31]. A time series itself is a sequence of data points indexed in time order, commonly encountered in diverse domains such as finance, healthcare, environmental science, and industrial processes [32]. The primary

objective of TSC is to build a model that can learn discriminative patterns or features from a collection of labeled time series, enabling it to accurately predict the class of new, unseen time series instances, and providing a powerful framework for automated detection tasks [31, 33].

Early TSC methods often relied on distance-based measures, with Dynamic Time Warping (DTW) being a prominent example, which calculates similarity between two temporal sequences that may vary in time or speed [34]. Other approaches focused on extracting statistical or structural features from the time series (e.g., mean, variance, spectral properties, shapelets) and then applying standard static classification algorithms [33, 35]. While effective in certain contexts, these methods can sometimes struggle with the high dimensionality, inherent noise, and complex temporal dependencies often present in real-world time series data.

More recently, deep learning architectures have demonstrated state-of-the-art performance on many TSC benchmarks [36]. Prominent among these are variants of Convolutional Neural Networks (CNNs) and Residual Networks (ResNet), adapted for time series, and have achieved remarkable success by effectively capturing local and global patterns [37]. Methods like InceptionTime also utilizes modules with multiple convolutional filter sizes to capture features at different temporal scales [38]. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) units and Gated Recurrent Units (GRUs), are inherently suited for modeling temporal dependencies and have also been widely applied, though sometimes at a higher computational cost [39]. Furthermore, Transformer-based models, leveraging self-attention mechanisms, have shown significant promise in capturing long-range dependencies and complex relationships within time series data [40]. Other approaches also includes ensemble methods, which combine several heterogeneous classifiers into a large meta-ensemble [41, 42, 43]. However, training these models can be computationally intensive, require large datasets, which poses limits in a sparse data scenario.

Alongside the advancements in deep learning, a distinct and highly competitive paradigm has gained prominence, focusing on transform-based feature engineering coupled with efficient classification. This approach decouples the often complex task of feature extraction from the subsequent classification step, allowing for specialized, powerful feature generation techniques to be paired with simpler, faster classifiers. Early and comprehensive exemplars of this philosophy include toolkits like hctsa (Highly Comparative Time-Series Analysis) [44] and catch22 [45]. hctsa facilitates the extraction of thousands of diverse time-series features derived

from a wide array of scientific literature, enabling extensive data-driven exploration and the potential discovery of novel discriminative patterns [44]. In contrast, catch22 offers a more concise, curated set of 22 canonical time-series characteristics selected for their broad applicability and high discriminative power in general-purpose TSC tasks, aiming for robust performance with minimal redundancy [45]. As a state- of-the-art framework in TSC, catch22 shares conceptual similarities with detach-ROCKET in its approach to feature pruning. Both these methodologies transform raw time series into a rich feature space, thereby empowering relatively simple classifiers to perform effectively.

Another highly influential branch within this "transform-then-classify" paradigm emerged with ROCKET (RandOm Convolutional KErnel Transform) [14]. ROCKET demonstrated that convolving time series with a large number of randomly generated, diverse convolutional kernels and extracting simple summary statistics (typically the maximum value and proportion of positive values) could yield highly discriminative features. These features then enable robust linear classifiers, such as ridge regression, to achieve state-of-the-art accuracy with exceptional speed, bypassing the need for complex backpropagation and extensive hyperparameter tuning associated with many deep learning models [14, 46]. The success of ROCKET spurred further innovations: MiniROCKET significantly reduced computational overhead by optimizing kernel parameters and focusing primarily on the proportion of positive values feature [15], while MultiROCKET enhanced performance by incorporating a richer set of summary statistics from the convolved outputs and an improved kernel generation strategy [47]. Building upon this lineage, Detach-ROCKET further reinforces the modular nature of the approach by separating, or "detaching" the redundant random kernels from the classifier training stage [16]. This design preserves computational efficiency and enhances the flexibility of feature generation, thereby readily supporting ensemble methods such as Detach Ensemble [48], where diverse feature sets can be strategically combined. The collective strength of these methods lies in their computational efficiency, reduced reliance on massive datasets, and potential for greater interpretability, making them particularly well-suited for challenging biomedical signal analysis tasks.

2.4 Related works

The automated detection of epileptic seizures from EEG signals has been an active area of research for several decades, driven by the need to alleviate the

laborious and subjective process of manual EEG review by neurologists, and enable timely clinical intervention [9, 49].

Early approaches predominantly relied on extracting salient features from EEG segments in various domains: time, frequency, and time-frequency—followed by classification using conventional machine learning algorithms [10, 50]. Common features include statistical measures (e.g., variance, kurtosis), spectral power in different EEG bands (delta, theta, alpha, beta, gamma), wavelet coefficients, and entropy measures [51, 52]. Classifiers such as Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN), Random Forests (RF), and Artificial Neural Networks (ANNs) were then trained on these extracted features to distinguish ictal (seizure) from interictal (non-seizure) or normal EEG patterns [11, 52]. While these methods have shown success, their performance heavily depends on the quality and relevance of the handcrafted features, which often require domain expertise and may not generalize well across diverse patient populations or recording conditions [9, 49].

More recently, deep learning techniques have demonstrated the ability to automatically learn hierarchical feature representations directly from raw or minimally processed EEG data [12, 13, 53]. Convolutional Neural Networks (CNNs) have been widely adopted for their proficiency in capturing spatial and temporal patterns from EEG signals, often treating multichannel EEG as an image or a set of parallel time series [54, 55, 56]. Recurrent Neural Networks (RNNs), particularly LSTMs and GRUs, are well-suited for modeling the temporal dynamics and long-range dependencies in EEG sequences [57, 58]. Hybrid models combining CNNs for feature extraction and RNNs for sequence modeling (CNN-LSTM) have also shown promising results [59, 60]. Furthermore, Transformer models, with their self-attention mechanisms, are increasingly being explored for EEG analysis, including seizure detection, due to their capability to capture global dependencies in long sequences [53, 61, 62, 63, 64, 65, 66]. These deep learning approaches often achieve good performance but typically require substantial amounts of labeled data for training and can be computationally intensive [12, 13].

Some recent works also explore the application of ROCKET-based methodologies for EEG feature learning in seizure detection, aiming to combine the speed of random convolutional kernel transforms with effective classification [67, 68]. These approaches seek to provide a balance between high accuracy and computational efficiency, which is crucial for practical clinical applications, especially in resource-constrained environments or for real-time monitoring systems [69, 70].

Chapter 3

Method

3.1 Data Processing

3.1.1 Data

In this project, two datasets are employed for model training and evaluation. The first dataset is the TUH EEG Seizure Corpus [71] (TUSZ), which comprises recordings from a total of 459 subjects. Within this corpus, a subset of the EEG recordings contains annotated seizure events, while the remainder consists of background (BCKG) recordings included to balance the dataset and to better assess the system's false alarm performance. Here BCKG recordings are defined as those that do not exhibit any of the following patterns: spike and/or sharp waves, periodic lateralized epileptiform discharges, generalized periodic epileptiform discharges, eye movements, or artifacts. Each EEG recording consists of 19 channels corresponding to the standard 10–20 system scalp electrode montage (see Figure 3.1), with a sampling rate of 256 Hz. See the publication of Obeid et al. [72] for further details on data collection.

Another dataset we used is the Siena Scalp EEG Database [73]. This database contains EEG recordings from 14 patients, collected at the Unit of Neurology and Neurophysiology at the University of Siena. The participants include 9 males (ages 25–71) and 5 females (ages 20–58). Each subject was monitored using Video-EEG at a sampling rate of 256 Hz. The recordings include 19 EEG channels, based on the standard 10–20 system scalp electrode montage (see Figure 3.1). See the publication of Detti et al. [74] for further details on data collection.

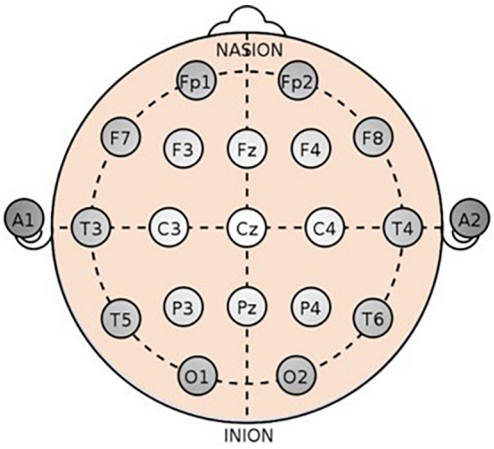


Figure 3.1: Electrode locations for EEG signals. From Shah, V. et al [71]

3.1.2 Data preprocessing

To support model training across multiple datasets, we convert each dataset to the Seizure Community Open-Source Research Evaluation (SzCORE) standardized format [75] for data and seizure annotations. This format is compatible with the Brain Imaging Data Structure (BIDS).

The model is trained on the large Temple University Seizure Corpus (TUSZ) dataset, which includes 1,134 recordings with seizures and 4,585 background recordings. All recordings are downsampled to 128 Hz. Each recording is segmented into non-overlapping 10-second epochs. These epochs are then categorized as seizure, interictal, or background (bckg). Seizure epochs are sampled from periods annotated as seizures. Interictal epochs are sampled from intervals between annotated seizure events. Background epochs are sampled exclusively from recordings labeled as BCKG. Seizure epochs are assigned the label 1 (seizure), while interictal and background epochs are assigned the label 0 (non-seizure).

3.2 Model

3.2.1 ROCKET and MINIROCKET

ROCKET (**R**and**O**m **C**onvolutional **KE**rnel **T**ransform) achieves state-of-the-art classification accuracy for time series classification while requiring only a fraction of the computational resources used by most existing methods [14].

It transforms time series using random convolutional kernels and uses the transformed features to train a linear classifier. The method was later reformulated as MINIROCKET (**MINI**mally **R**and**O**m **C**onvolutional **KE**rnel **T**ransform), which is up to 75 times faster than ROCKET on large datasets and is almost deterministic, while maintaining nearly the same accuracy [15]. MINIROCKET is significantly faster than other methods with similar accuracy and provides significantly better accuracy than methods with comparable computational efficiency.

3.2.2 Detach-ROCKET

Although ROCKET and MINIROCKET are efficient and computationally lightweight, many of the randomly generated features are redundant or non-informative, increasing the computational burden and may reduce the model's generalizability. Detach-ROCKET addresses this issue by introducing Sequential Feature Detachment (SFD), a method designed to identify and prune the non-essential features from ROCKET-based models [16]. In SFD, the transformed features are ranked according to their contribution to the model's decisions. At each iteration, a fixed proportion of the least informative features is discarded. Let $\mathbb{F}$ denote the complete set of features generated by ROCKET's random convolutional kernels. The subset of currently active features is represented by $\mathbb{S} \subseteq \mathbb{F}$, which is initially set to $\mathbb{F}$.

At each step t, a ridge classifier is trained on the active feature set $\mathbb{S}_t$ by solving the following optimization problem:

$$\hat{\theta}_t^{\text{ridge}} = \underset{\theta}{\text{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \theta_0 - \sum_{k \in \mathbb{S}_t} x_{ik}\theta_k \right)^2 + \lambda \sum_{k \in \mathbb{S}_t} \theta_k^2 \right\} \tag{3.1}$$

The training process produces a set of optimal coefficients $\hat{\theta}_t^{\text{ridge}} = \{\hat{\theta}_k\}$, where each coefficient is proportional to the contribution of a corresponding feature on the classifier's decision. The features are then ranked according to the absolute values of their coefficients. Let p denote the proportion of features to be removed. At each step, the lowest $100 \cdot p\%$ of ranked features discarded, and the remaining $100 \cdot (1-p)\%$ are retained to form the updated feature set $\mathbb{S}_{t+1}$. Since p controls the trade-off between computational cost and model accuracy, it is set to 0.05 to ensure a conservative yet computationally affordable pruning procedure, meaning that 5% of the features are removed at each step.

To determine the optimal number of features to finally retain in our dataset, we solve the following optimization problem:

$$Q_c = \underset{q}{argmax}\ f_c(q) = \underset{q}{argmax}\ \{\alpha(q) + c \cdot q\}\,. \tag{3.2}$$

In this equation, Q_c represents the accuracy curve obtained by evaluating the model at each pruning step, q represents the proportion of pruned features, and $\alpha(q)$ the accuracy of the pruned model on the validation set. The hyperparameter c is the weighting factor between accuracy and data size, in which a smaller c value favors accuracy and larger c value favors data size. We set $c = 0.1$, as this achieves the optimal performance according to Fig.4 in [16].

3.2.3 Detach Ensemble

Given the high dimensionality of multivariate time series such as the 19-channel EEG data in this project, relying on a single set of randomly generated kernels may inadequately capture the complex spatio-temporal patterns and inter-channel relationships critical for seizure detection. To address this limitation and improve model generalization and robustness, an ensemble approach based on Detach-ROCKET is employed. The Detach Ensemble involves training N independent Detach-ROCKET models. For each model, a subset of the training data is used to determine the optimal pruning size, and the model is then pruned using SFD. Each pruned model is then assigned a weight based on its performance on the training set [48].

For classifying a given input instance, predictions from each individual model are aggregated via a weighted average, using the predetermined model weights, to yield a final ensemble probability score. This probability is then thresholded (commonly at 0.5) to produce the definitive classification label (seizure or non-seizure).

In our implementation, we use an ensemble of $N = 10$ Detach-MiniROCKET models. This number was chosen as a balance between seeking improved performance through ensemble diversity and managing computational resources, as training significantly more models becomes resource-intensive. Each of the 10 model is trained independently using the procedure described in Section 3.2.2, including the application of SFD with $p = 0.05$ and the feature selection criterion $c = 0.1$.

3.2.4 catch22

To establish a robust, interpretable, and accessible baseline for our models, we selected the catch22 (CAnonical Time-series CHaracteristics) feature set [45] as our benchmark. catch22 provides a curated collection of 22 highly discriminative time-series features, systematically selected from the larger hctsa library to maximize performance while minimizing redundancy. This focus on feature selection makes it a relevant state-of-the-art comparison for our Detach-ROCKET methodology.

3.3 Evaluation of Models

The performance of the trained Detach-ROCKET and Detach Ensemble model are assessed through two complementary evaluation frameworks: epoch-wise and event-wise analysis.

Firstly, an epoch-wise evaluation is conducted based on the 10-second non-overlapping segments described in Section 3.1.2. Standard classification metrics including accuracy, sensitivity, precision, and the F1-score are computed on the test set. The confusion matrix is also examined to understand the distribution of classification errors. This epoch-based assessment provides views of the model's ability to correctly classify individual epochs.

However, recognizing that clinical assessment of epilepsy monitoring often focuses on the detection of seizure episodes (events) rather than isolated time epochs, an event-wise evaluation is performed. During the evaluation procedure, event-based sensitivity, precision, and F1-score are calculated base on the open-source `timescoring` library [76].

3.3.1 Evaluation metrics

To evaluate the performance both epoch-wise and event-wise, we first define and explain the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) in this study: True Positive (TP): An instance correctly identified as belonging to the positive class (seizure); True Negative (TN): An instance correctly identified as belonging to the negative class (non-seizure);False Positive (FP): An instance incorrectly identified as belonging to the positive class when it belongs to the negative class (a non-seizure instance classified as seizure);False Negative (FN): An instance incorrectly identified as belonging to the negative class when it belongs to the positive class (a seizure instance classified as non-seizure).

For epoch-wise evaluation, a TP is a correctly classified seizure epoch, an FP is a non-seizure epoch classified as seizure, etc. For event-wise evaluation, the following parameters were configured to combine the epochs to seizure events, as suggested by the 2025 Seizure Detection Challenge:

- Minimum Overlap: Any temporal overlap, however brief, between a reference event and a hypothesis event is sufficient to consider it a potential match. This setting maximizes sensitivity to detecting any part of a seizure.
- Pre-ictal Tolerance: A hypothesis event starting up to 30 seconds before the onset of a reference event can still be considered a detection of that event.
- Post-ictal Tolerance: A hypothesis event ending up to 60 seconds after the end of a reference event can still be considered part of the detection of that event.
- Minimum Duration: Reference or hypothesis events separated by less than 90 seconds are merged into a single, longer event before scoring. This duration corresponds to the sum of the pre- and post-ictal tolerances, preventing closely spaced detections from being penalized multiple times.

Figure 3.2 illustrates the epoch and event definitions used in our evaluation framework. In this configuration, a reference event refers to a labeled seizure event, a hypothesis event refers to a model predicted event. Based on these criteria, an event-based TP occurs when a reference seizure event is correctly matched with one or more hypothesis seizure events according to the overlap and tolerance rules. An event-based FP corresponds to a hypothesis seizure event that does not match any reference event. An event-based FN represents a reference seizure event that is not matched by any hypothesis event. For the specific application of seizure detection, event-wise metrics are considered more clinically meaningful than epoch-wise metrics. Consequently, while both will be reported, event-wise performance will be the primary focus for evaluating the model's practical utility.

Based on these setting, we define the accuracy, sensitivity, precision, f1-score and false alarm rate as following:

- Accuracy: the proportion of total instances that were correctly classified.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

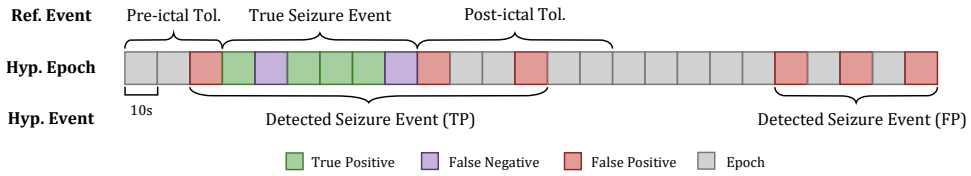


Figure 3.2: Epoch-wise and event-wise evaluation framework for seizure detection. Discrete epoch-level model predictions (Hyp. Epoch) are aggregated into continuous seizure events (Hyp. Event) following 2025 Seizure Detection Challenge aggregation rules. Detected events are then matched against ground truth reference events (Ref. Event) within specified tolerance windows to determine final classification outcomes.

- Sensitivity: the proportion of true positive instances that were correctly identified by the model.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3.4}$$

- Precision: the proportion of instances predicted as positive that were actually positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.5}$$

- F1 Score: the harmonic mean of Precision and Sensitivity, providing a single metric that balances both concerns.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3.6}$$

- False positive rate (FPR): the number of false positives per 24 hours. A low FPR is crucial for clinical usability, as frequent false alarms can reduce clinician trust in the system.

$$\text{FPR} = \frac{FP}{FP + TN} \tag{3.7}$$

3.4 Channel Relevance Estimation

To estimate the relevance of individual EEG channels within our Detach-ROCKET Ensemble, we adopt the methodology proposed by Solana et

al. [48]. This process first assesses channel relevance on each Detach-MINIROCKET model:

1. **Kernel Selection:** Sequential Feature Detachment (SFD) identifies and selects the most relevant kernels within the MiniROCKET model.
2. **Channel Retrieval:** For each kernel selected by SFD, the specific EEG channels it processed are retrieved.
3. **Importance Weighting:** Each retrieved channel receives an importance score proportional to its kernel's weight (θ_i). It is then divided by the number of channels in that same kernel.
4. **Relevance Aggregation:** The weighted importance scores for all channels are summed across the selected kernels. These sums are then normalized to create a relative channel relevance histogram.

The final ensemble relevance for each channel is then determined by taking the median of relevancies from base models and normalizing across all channels.

Chapter 4

Results and Analysis

4.1 Experiments on TUSZ Corpus

4.1.1 Training and test with Detach-MINIROCKET and Ensemble

In this project, we compared two models, Detach-ROCKET and Detach Ensemble, on the TUSZ dataset, using catch22 as a benchmark.

Due to hardware constraints and the large size of the dataset, we randomly select 50% of the available subjects to construct the working dataset. This subset is then divided into training and test sets based on subject identifiers to ensure that the evaluation reflects the model's generalizability across individuals. We allocate 80% of the sampled subjects to the training set and the remaining 20% to the test set. Each subject includes multiple sessions and runs, with each run containing several recordings labeled as either seizure or BCKG (See Figure 4.1 for an overview). Figure 4.2 displays a representative segment of a multi-channel EEG recording, showing the complex electrographic characteristics of its pre-ictal, ictal, and post-ictal phases. Notably, a portion of the subjects in the dataset contribute exclusively BCKG recordings, without any associated seizure activity.

To accommodate memory limitations and promote a balanced and diverse training set, after segmenting the recordings in the training set into non-overlapping 10-second epochs, we include all available seizure epochs, randomly select an equal number of interictal epochs, and randomly sample half that number of background epochs for model training.

Models' performances were initially assessed using epoch-wise accuracy. Table 4.1 presents the accuracy performance of both models, showing similar

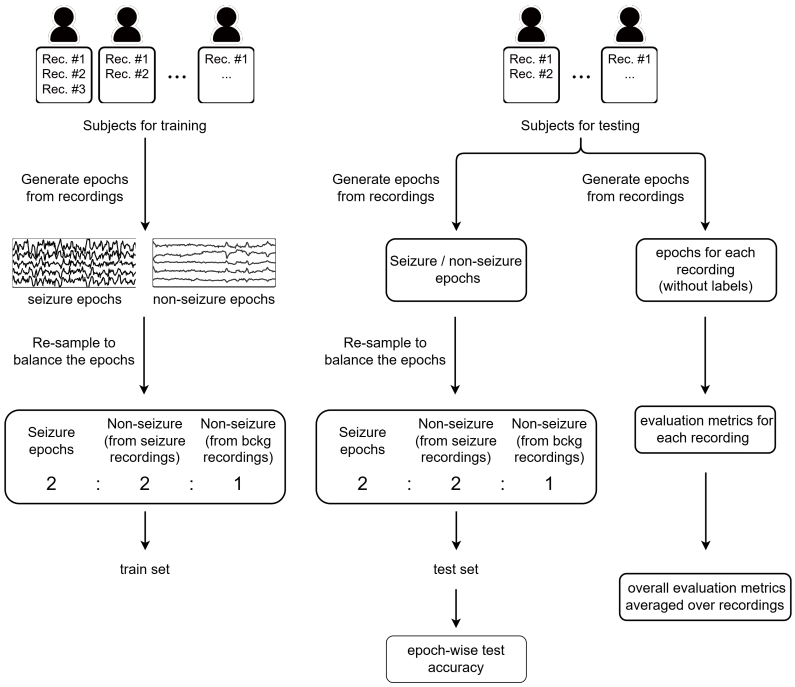


Figure 4.1: Overview of the experiment pipeline, detailing the data preparation steps, the training pipeline with balanced epoch sampling, and the evaluation pipeline leading to both epoch-wise and event-wise performance metrics.

test accuracy across approaches. The models are then evaluated on a dataset comprising the same subjects included in the test set during model training, while using all the epochs produced by these subjects without selecting a subset of them. Note that the evaluation dataset, though comprising the same subjects as test set, is however unbalanced with more non-seizure epochs than seizure epochs. Performance was evaluated using sensitivity, precision, F1-score, and False Positive Rate (FPR per 24 hours), computed at the recording level for both epoch-wise and event-wise analyses. In this study, event-wise metrics are considered more clinically meaningful than epoch-wise metrics. While epoch-wise metrics provide valuable insights into model behavior, event-wise metrics, particularly the event-wise F1 score, serve as the primary indicators for model evaluation due to their direct clinical relevance.

Table 4.2 presents performance metrics for a single Detach-MINIROCKET model (10,000 features) and a 10-model Detach-MINIROCKET ensemble (each base model using 10,000 features). Metrics were computed for each evaluation recording and then averaged across all recordings. We report the median values and min-max ranges derived from three identical

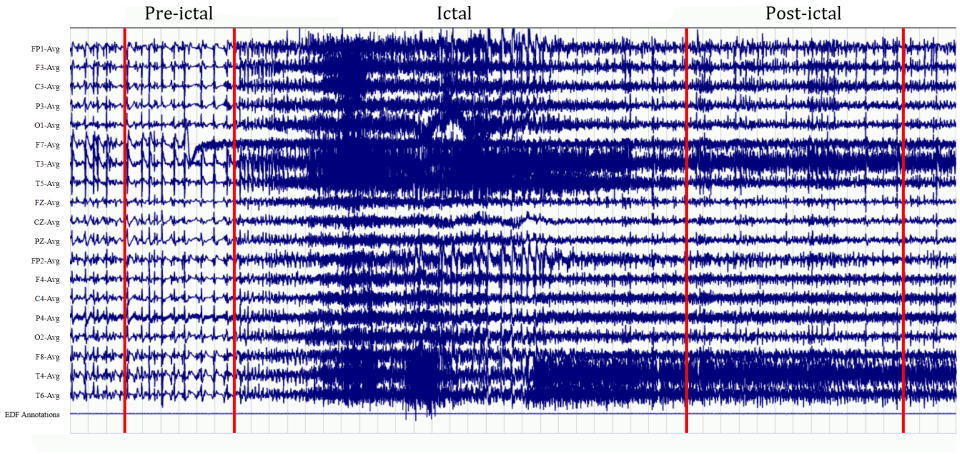
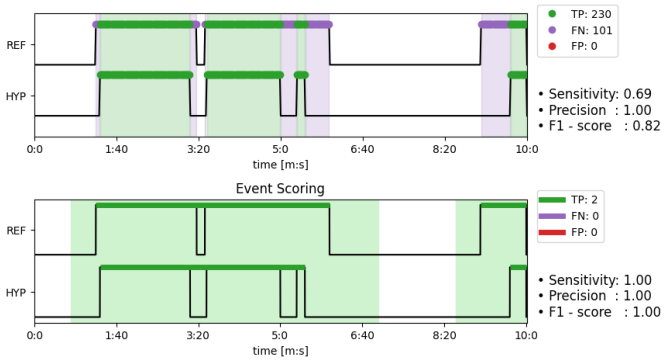


Figure 4.2: A representative EEG segment showing a clinically annotated seizure event (labeled 'Ictal') from TUSZ dataset. The 'Pre-ictal' and 'Post-ictal' labels illustrate the tolerance windows (30s and 60s respectively) used in our event-scoring methodology. This example, with its complex signal characteristics, highlights the challenge of accurately isolating the true ictal pattern even with defined evaluation tolerances.

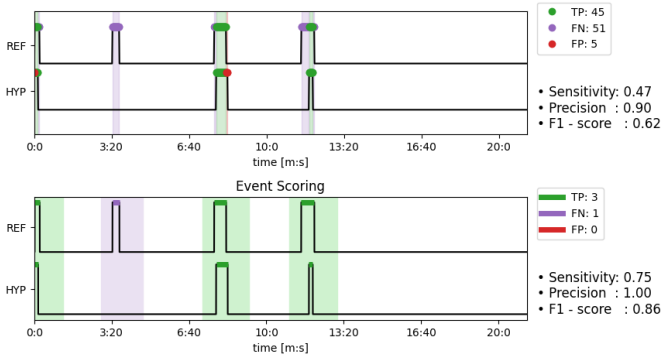
random data partitions. Figure 4.3 illustrates the evaluation framework on representative recordings containing both seizure and background EEG patterns. It demonstrates how model predictions (HYP) are compared against a ground truth reference (REF) to determine True Positives (TP), False Negatives (FN), and False Positives (FP) at both the epoch-wise level (top panel) and the clinically relevant event-wise level (bottom panel).

The high event-wise F1 scores achieved by both models (Table 4.2) demonstrate strong generalization capabilities across different individuals. However, the single Detach-MINIROCKET model shows a slightly (4%) lower event-wise F1 score and a 9% lower epoch-wise F1 score compared to the ensemble, suggesting that a single Detach-ROCKET model with 10,000 features may have limited capacity to adequately sample the feature space and capture meaningful multichannel information. Although the Detach Ensemble incurs higher computational costs compared to Detach-MINIROCKET (see Section 4.4 for a quantitative comparison), it offers significant advantages beyond performance by providing label probabilities and channel relevance estimation, which are particularly valuable features for seizure detection applications.

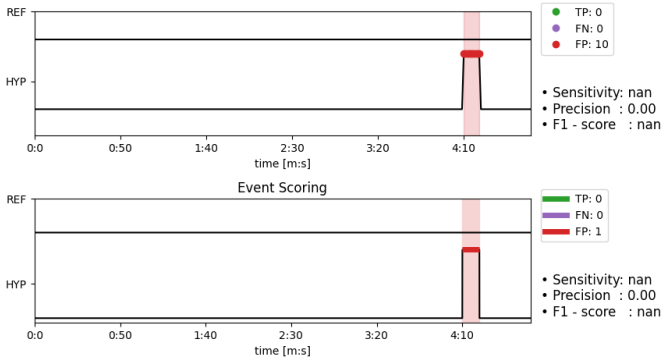
To further assess the model's generalization capabilities across individual subjects, subject-wise performance metrics were computed on the test set to



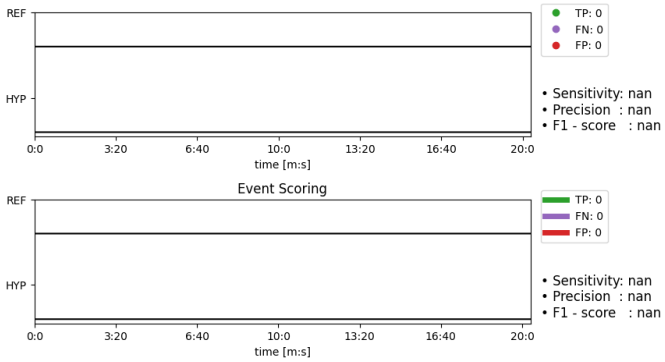
(a) Correctly predicts all seizure events (TP) with no FP.



(b) Fails to predict an existing seizure event (FN).



(c) Incorrectly predicts a seizure when no seizure is present (FP) in a BCKG recording.



(d) Correctly predicts the absence of seizure (TN) in a BCKG recording.

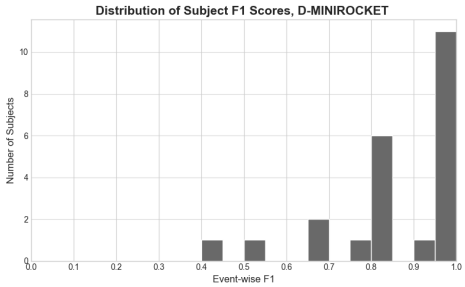
Figure 4.3: Example of epoch-wise(top panel) and event-wise(bottom panel) scoring for analyzing the performance on an evaluation recording.

investigate the extent of inter-subject variability in prediction performance. Figure 4.4 illustrates the distribution of event-wise F1 scores and False Positive Rates(per 24h) across individual subjects from a representative evaluation, revealing considerable inter-subject variability. For both models, a significant proportion of subjects obtained high F1 scores (above 0.8). In fact, for both models (Figures 4.4c and 4.4a), 10 of the 11 subjects in the top performance bin (0.95-1.0) achieved an F1 score of 1.0. FPR values also showed considerable variation across subjects, with a large proportion of subjects demonstrating low false alarm rates(below 20). Notably, within the lowest FPR bin (0-20 FP/24h), the vast majority of subjects had an FPR of zero, with 11 out of 13 for the Detach Ensemble (Figure 4.4d) and 11 out of 12 for the single Detach-MINIROCKET model (Figure 4.4b. Despite the observed inter-subject performance variation, these distributions suggest that both models generalize effectively for a majority of individuals. It is important to note that the F1 score distribution presented in Figure 4.4 excludes subjects who contributed exclusively background (BCKG) recordings to the test set (i.e., had no annotated seizure events). For these subjects, True Positives (TP) and False Negatives (FN) are inherently zero, so the F1 score cannot be calculated.

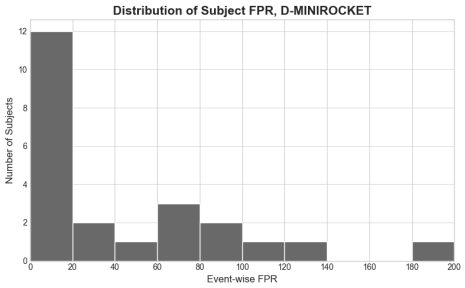
4.1.2 Benchmark

To establish a robust, interpretable, and accessible baseline for our models, we selected catch22 [45] as our benchmark. catch22 offers a concise set of 22 highly discriminative and computationally inexpensive time-series features, distilled from the much larger hctsa feature set [44]. Its design emphasizes generalizability and interpretability across diverse time series domains [31, 45]. Furthermore, the widespread availability and easy implementation of catch22 contribute to its utility as a practical and reproducible benchmark for evaluating the performance and computational costs of more complex or specialized algorithms in TSC tasks.

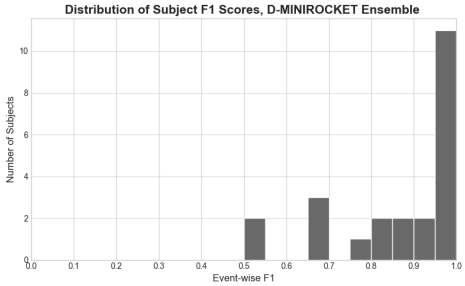
Table 4.1 and Table 4.2 summarize a comparative performance with the benchmark model, averaged over three identical random splits, each using 50% of the total available subjects. When evaluated on the TUSZ test set, all models demonstrated similar median event-wise F1 scores. The Detach Ensemble showed slightly better stability with a performance range range of 0.87-0.91, compared to the 0.84–0.88 range of both the single Detach-MINIROCKET and catch22. At the epoch level, the Detach Ensemble and catch22 were more effective, outperforming the single Detach-MINIROCKET by 9% and 5% respectively.



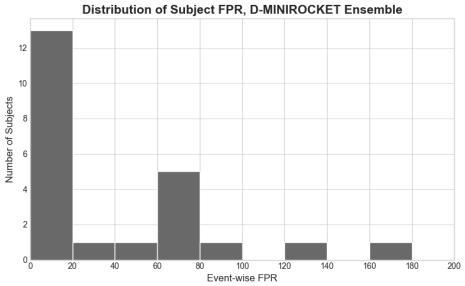
(a) Subject-wise event F1 score distribution for the Detach-MINIROCKET on the TUSZ test set.



(b) Subject-wise event FPR distribution for the Detach-MINIROCKET on the TUSZ test set.



(c) Subject-wise event F1 score distribution for the Detach Ensemble on the TUSZ test set.



(d) Subject-wise event FPR distribution for the Detach Ensemble model on the TUSZ test set.

Figure 4.4: Distribution of subject-wise performance metrics on the TUSZ test set for different model configurations.

Table 4.1: Comparison of epoch-wise training and test accuracies for the single Detach-MINIROCKET (D-MINIROCKET) model, Detach-MINIROCKET Ensemble (D-MINIROCKET Ens.) and catch22. Results show median values with corresponding minimum–maximum ranges from three runs each model.

	Train Acc.	**Test Acc.**
D-MINIROCKET	0.97	0.74
Range	*(0.96, 0.97)*	*(0.72, 0.76)*
D-MINIROCKET Ens.	0.96	0.73
Range	*(0.96, 0.97)*	*(0.71, 0.85)*
catch22	1.00	0.76
Range	*1.00*	*(0.65, 0.87)*

Table 4.2: Comparison of model performance for single Detach-MINIROCKET (D-MINI), 10-model Detach-MINIROCKET ensembles(D-MINI Ens.) and catch22, using 50% of the subjects of the TUSZ dataset. Results show median values with corresponding minimum–maximum ranges from three runs each model.

Config.	**Level**	**Sens.**	**Prec.**	**F1**	**FPR/24h**
D-MINI	Epoch	0.72	0.60	0.48	16 428
	Range	*(0.49, 0.72)*	*(0.46, 0.60)*	*(0.42, 0.58)*	*(10308, 20198)*
	Event	0.95	0.81	0.85	43
	Range	*(0.93, 0.98)*	*(0.80, 0.84)*	*(0.84, 0.88)*	*(35, 49)*
D-MINI Ens.	Epoch	0.73	0.63	0.57	17 298
	Range	*(0.54, 0.81)*	*(0.54, 0.70)*	*(0.49, 0.57)*	*(9080, 18829)*
	Event	0.97	0.87	0.89	29
	Range	*(0.91, 0.98)*	*(0.86, 0.89)*	*(0.87, 0.91)*	*(25, 35)*
catch22	Epoch	0.58	0.64	0.53	9 756
	Range	*(0.44, 0.62)*	*(0.62, 0.70)*	*(0.41, 0.56)*	*(9660, 10392)*
	Event	0.91	0.86	0.88	31
	Range	*(0.89, 0.95)*	*(0.86, 0.88)*	*(0.84, 0.88)*	*(25, 34)*

4.2 Sensitivity Analysis

The performance of a single Detach-MINIROCKET model, and consequently its sensitivity to hyperparameter changes, can be significantly influenced by the stochastic nature of its random convolutional kernel generation [14, 15]. While this inherent randomness is integral to the efficiency of the ROCKET family, it makes it more difficult to isolate the specific impact of individual hyperparameter adjustments in limited model instances, as it can have a large variation due to randomness rather than solely the hyperparameter change. In contrast, Detach-Ensemble, by aggregating multiple Detach-MINIROCKET instances, reduces such stochastic effects, leading to more stable and representative performance characteristics. Given the foundational homogeneity between Detach Ensemble and its constituent Detach-MINIROCKET models (as the ensemble is built from them), our detailed sensitivity analysis was primarily focused on the Detach-Ensemble configuration.

We explored the effect of key parameter choices on the Detach-Ensemble's test accuracy, epoch-wise and event-wise F1 score. The key parameters investigated were:

1. number of models (N)
2. epoch duration
3. training data size

Table 4.3 presents the results for Detach Ensemble models configured with different N and epoch durations trained on the same data split of the TUSZ dataset. Both epoch-wise and event-wise F1 scores demonstrated considerable robustness to changes in epoch duration ranging from 6 to 20 seconds. Similarly, varying the number of ensemble models (N) between 5 and 20 did not substantially affect the F1 scores. Epoch-wise test accuracy is observed to be more sensitive to hyperparameter changes. Optimal accuracy was achieved with $N = 10$ models and 6-second epochs, while the lowest occurred with $N = 20$ models and 10-second epochs. This implies 6-second epochs might offer a more effective temporal window for seizure discrimination in this dataset. Given that event-wise performance did not improve when varying number of models used in the ensemble, our selection of $N = 10$ is therefore an efficient choice.

Table 4.3: Performance evaluation of the Detach-Ensemble model under varying hyperparameter configurations.N denotes the number of models in the ensemble.

N	**Epoch Duration**	**Test Acc.**	**Epoch F1**	**Event F1**
10	10	0.80	0.62	0.90
10	6	0.85	0.64	0.89
10	20	0.81	0.62	0.89
5	10	0.84	0.63	0.88
20	10	0.78	0.62	0.89

The impact of training data size on model performance is shown in Figure 4.5. A notable deterioration in performance is observed when the proportion of training data used falls below 30% of the total dataset. Beyond this point, the model achieves a relatively stable performance. This finding supports our decision to use 50% of the available subjects for training, as it offers a conservative, resource-efficient approach while ensuring robust model performance. The result also suggests model's ability for effective generalization even when trained on limited dataset sizes.

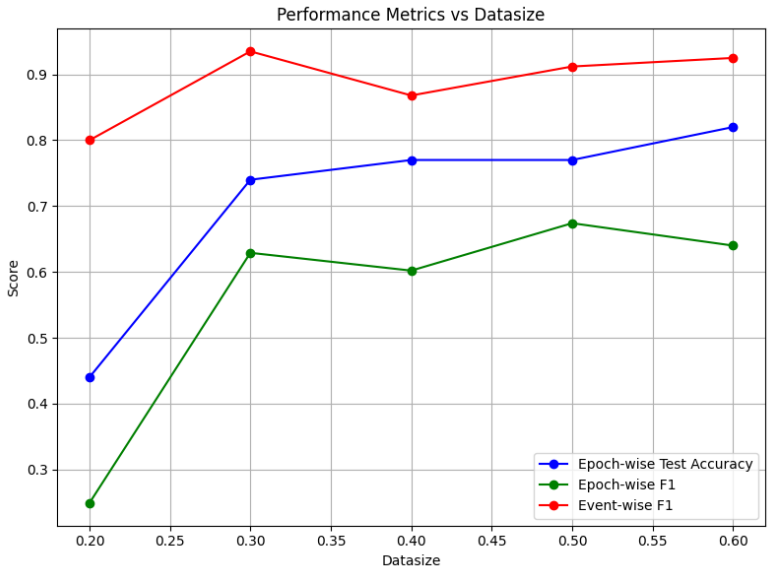


Figure 4.5: Performance of the Detach Ensemble model on varying training subjects proportions. The plot shows epoch-wise test accuracy, epoch-wise F1 score, and event-wise F1 score.

4.3 Cross-Dataset Evaluation

To evaluate the models' generalization capabilities on unseen data, we assess model performance on the Siena Scalp EEG Database. This independent dataset provides an opportunity to test the model's robustness to variations in patient populations, recording equipment, and potential differences in EEG characteristics. Table 4.4 presents the overall performance of our models when evaluated on the entire Siena dataset, compared with catch22. The results demonstrate that the Detach-Ensemble model achieved superior performance compared to both Detach-MINIROCKET and catch22. Specifically, the event-wise F1 scores were 7% and 6% higher, and the epoch-wise F1 scores were 6% and 9% higher, respectively. The ensemble model also resulted in a lower event-wise False Positive Rate per 24 hours. Detach-MINIROCKET and catch22 achieved similar performance.

To further understand models' cross-dataset generalization capabilities, we refer to the publicly reported outcomes of the Seizure Detection Challenge 2025. In this challenge, proposed algorithms were trained on publicly available datasets and then evaluated on a private EEG dataset (Dianalund Scalp EEG dataset), with rankings primarily determined by their event-wise

F1 score. Adhering to the same epoch-to-event aggregation rules employed in the challenge, our Detach-Ensemble model achieved an event-wise F1 score of 0.39. This performance is highly competitive with the top-ranked algorithms in the challenge, which reported event-wise F1 scores of 0.43, 0.36, and 0.34. Given that the Siena dataset represents an entirely different data distribution from the training set (TUSZ), the performance of our model indicates a strong capacity for generalization and robust seizure detection across diverse datasets.

Table 4.4: Cross-dataset performance comparison of Detach-MINIROCKET, Detach-MINIROCKET Ensemble, and catch22 on the Siena dataset. Results show median values with corresponding minimum–maximum ranges from three runs each model.

	Epoch F1	**Event F1**	**Event FPR**
D-ROCKET	0.28	0.32	81
Range	*(0.22, 0.32)*	*(0.28, 0.42)*	*(58, 93)*
D-ROCKET Ens.	0.34	0.39	63
Range	*(0.32, 0.37)*	*(0.34, 0.45)*	*(51, 76)*
catch22	0.25	0.33	91
Range	*(0.24, 0.28)*	*(0.27, 0.37)*	*(82, 116)*

4.4 Computational Efficiency of Model Inference

The computational efficiency of the models, particularly their prediction speed, is a critical factor for practical deployment, especially in real-time seizure detection scenarios. To evaluate this, prediction times were measured on an Intel(R) Xeon(R) Platinum 8480C processor for a representative EEG recording of 1208 seconds in duration. The results are summarized in Table 4.5.

As indicated in Table 4.5, the single Detach-MINIROCKET model processed the entire recording in just 0.74 seconds, which is approximately 17.1 times faster than the catch22 benchmark. The Detach-MINIROCKET Ensemble, while inherently introducing a higher computational cost due to the aggregation of multiple individual models, still completed the prediction in 6.88 seconds. This remains notably faster than catch22, by a factor of approximately 1.8. These findings highlight the significant computational

advantages of the ROCKET-based feature transformation approach employed by our models. In particular, the rapid inference capability of the single Detach-MINIROCKET model demonstrates strong potential for applications requiring near real-time processing and decision-making. While the ensemble model requires increased prediction time, it provides enhanced predictive performance than Detach-MINIROCKET and is therefore still a competitive choice.

Table 4.5: Model prediction time (seconds) on a 1208s EEG recording.

D-MINIROCKET	**D-MINIROCKET Ens.**	**Catch22**
0.74	6.88	12.68

4.5 Channel Relevance

To investigate the relative importance of different EEG channels for seizure detection within our framework, we employed the channel relevance estimation method proposed by Solana et al. [48]. This method was applied to each of the well-performing Detach-MINIROCKET Ensemble models developed during our training phase. Figure 4.6 illustrates the mean and standard deviation of the estimated channel relevance values, aggregated across these selected ensemble models.

Observations from Figure 4.6 indicate that channels Fp1 and C4 exhibit notably high mean relevance values compared to other channels. However, these channels also display relatively large standard deviations. Conversely, channel F8 presents with a markedly low mean relevance value and a comparatively small standard deviation, indicating its consistently minor role in seizure detection for our models on the TUSZ dataset.

Further analysis involved ranking channels from most to least relevant within each individual ensemble model. This ranking revealed that channels Fp1, C4, and O1 were frequently identified among the top four most relevant channels (appearing 4 out of 7 times for each of these channels across the evaluated models). In contrast, channel F8 was most frequently found among the top four least relevant channels (6 out of 7 model instances), with channel T3 also appearing frequently in this less relevant group (4 out of 7 model instances).

The consistent findings combining these two analytical approaches suggest that channels Fp1 and C4 are likely more critical for our seizure detection

task on the TUSZ dataset, while channel F8 appears to be consistently less informative. In practical scenarios where computational resources are constrained, or where feature selection is desired to emphasize inter-channel relationships or reduce model complexity, these findings suggest that prioritizing channels such as Fp1, C4, and potentially O1, while deprioritizing F8 and possibly T3, could be a viable strategy.

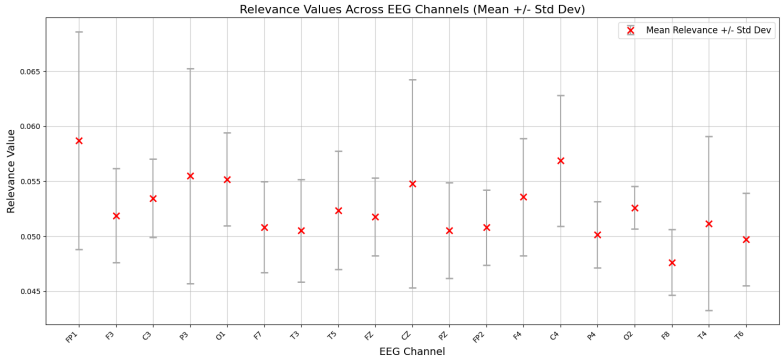


Figure 4.6: Mean and standard deviation of estimated channel relevance values across selected Detach-MINIROCKET Ensemble models. Higher mean values suggest greater overall importance, while standard deviation indicates variability across models.

Chapter 5

Discussion

This thesis addressed the challenge of developing a seizure detection algorithm that is both computationally efficient and robustly generalizable across diverse patient data. We successfully implemented and evaluated the Detach Ensemble model, which outperformed both the single Detach-MINIROCKET model and the catch22 benchmark in balancing performance and inference speed. On the TUSZ dataset, the model reached a high event-wise F1-score of 0.89 with impressive inference speed. Its generalization ability was further confirmed on the Siena dataset, where it achieved an F1 score of 0.39, comparable to top models from the 2025 Seizure Detection Challenge. These results establish the Detach Ensemble framework as a strong foundation for this task and set the stage for further discussion of our findings, limitations, and broader implications.

5.1 Benchmark selection

We chose catch22 as our benchmark because it provides a general and universal solution for time series classification (TSC) tasks. As a state-of-the-art framework in TSC, catch22 shares conceptual similarities with detach-ROCKET in its approach to feature pruning. The benchmark is easily accessible, well-documented, and has demonstrated effectiveness across numerous TSC applications. Notably, EEG datasets are included in the calibration set of time series classification tasks that catch22 employs, which provides some relevance to our seizure detection domain. However, it is important to acknowledge that catch22 is not specifically designed as a seizure detection method, nor is it tailored for the unique characteristics and requirements of seizure detection tasks. This limitation indicates that direct

comparisons may not constitute entirely fair evaluations, as the benchmark was not optimized for the specific challenges inherent in seizure detection, such as the temporal dynamics and the importance of minimizing false negatives in clinical applications.

It is important to point out that our model, Detach-ROCKET and Detach Ensemble, are not adjusted either for specific challenges in seizure detection. From this perspective, it is reasonable to conduct such a comparison. Importantly, the computational efficiency of catch22 features serves as a baseline for understanding the relative performance of our models in terms of efficiency. It also helps establish whether the additional complexity of our ensemble model yields meaningful improvements while not sacrificing efficiency excessively.

5.2 Subject Variations

A key observation from our evaluation is the substantial inter-subject variability in model performance, as illustrated by the subject-wise scores on the TUSZ dataset (Figure 4.4). This heterogeneity is not unexpected and reflects the inherent complexities of EEG-based seizure detection. Several factors likely contribute to these observed variations:

5.2.1 Variations in Data Composition

The composition of data available for each subject in the TUSZ corpus introduces a source of variability. Some subjects exclusively contribute background (BCKG) recordings, meaning the model's performance for these individuals is solely based on its ability to correctly identify non-seizure states and avoid false positives. Also, the number and length of seizures and the total duration of recordings can vary a lot between subjects with seizure events.

5.2.2 Inherent Seizure Heterogeneity

Epileptic seizures are highly heterogeneous phenomena, varying significantly not only between individuals but also potentially within the same individual over time. Subjects may exhibit different seizure types (e.g., focal vs. generalized, varying electrographic signatures) or present with idiosyncratic EEG patterns during ictal events. The Detach-MINIROCKET features, while robust, are learned from a diverse pool of seizures in the training set. However, if a particular subject's seizures show highly unique or

underrepresented electrographic characteristics compared to the dominant patterns in the training data, the model may struggle to generalize effectively to that specific individual.

5.2.3 Potential Variations in EEG Acquisition

While both the TUSZ and Siena datasets adhere to the standard 10-20 electrode placement system, the practical application of scalp electrodes can introduce subtle yet impactful deviations between subjects. Minor discrepancies in the precise anatomical landmarking for electrode positions can lead to considerable differences in the underlying cortical sources captured by each channel [77]. Therefore, the spatio-temporal patterns learned by the Detach-MINIROCKET model might be expressed differently across subjects due to placement variations. Additionally, inherent inter-subject differences in scalp thickness, skull conductivity, and electrode-scalp impedance further modulate the recorded EEG signals [78]. These variations can persist even after careful preprocessing and would possibly influence model performance on a subject basis.

5.3 Limitations

5.3.1 Computational Resource Constraints and Methodological Limitations

While the Detach-ROCKET and Detach-MINIROCKET models are inherently efficient and computationally lightweight compared to deep learning alternatives, our study was significantly constrained by the large scale of the EEG datasets and the associated memory requirements rather than the computational complexity of the models themselves. The TUSZ dataset, with its thousands of multi-channel EEG recordings, presents substantial challenges in terms of data loading, preprocessing, feature transformation storage, and overall memory management, particularly when operating within a resource-constrained environment. The ensemble methodology, while computationally efficient in terms of individual model training, amplifies the memory requirements due to the need to store and manage multiple model instances and their associated feature transformations.

Due to this limitation, we were constrained to use 50% of the available subjects from the TUSZ dataset for training and testing. While the models demonstrated sufficient performance on the test set(See Figure 4.5), this

reduced dataset size could potentially limit their generalizability to unseen data. Furthermore, these memory limitations restricted our ability to conduct a more extensive exploration of the hyperparameter space, which may in turn affect our thorough understanding of the effects these parameters have on model performance and behavior.

5.3.2 Impact of Randomness on Model Performance and Reproducibility

The performance variations observed across different experimental runs (See Table 4.2)in our study highlight a critical limitation inherent to the ROCKET-based methodologies: the influence of multiple sources of randomness on model performance and result reproducibility. First, the random partitioning of subjects into training and test sets introduces variability in the data distribution characteristics between different experimental runs. Given the substantial inter-subject variability discussed in Section 5.2, different subject splits can lead to training and test sets with markedly different difficulty levels and distributional properties. Second, the core ROCKET methodology relies on the generation of numerous random convolutional kernels with stochastically determined parameters including kernel length, weights, bias, dilation, and padding [14, 15]. The realization of these random kernels fundamentally determines the feature space representation and, consequently, the model's capacity to capture relevant temporal patterns for seizure detection.

These randomness sources are shown as substantial variation in reported performance metrics across different experimental runs, even when using identical hyperparameters and methodological configurations. Hyperparameter optimization and sensitivity analysis becomes less reliable when performance estimates are subject to high variance, especially with limited experimantal runs. Additionally, the comparison between different methodological approaches may be confounded by random variation rather than reflecting true algorithmic differences.

5.3.3 False Positive Rate

Our ensemble model exhibits 29 FP/24h in the TUSZ evaluation set and 63 FP/24h in Siena set, as shown in table 4.2 and table 4.4. These rates are too high for practical usage. Recent work by Ingolfsson et al. achieved substantially lower false positive rates in wearable EEG seizure detection,

reporting 0.51 FP/h on the PEDESITE dataset and 0.65 FP/h on the CHB-MIT dataset [79], equivalent to approximately 12.2 and 15.6 false positives per 24 hours, respectively. This represents a significant improvement over our current performance, highlighting false positive reduction as a critical area for future development.

5.4 Ethics and Sustainability

5.4.1 Ethical Concerns

The ethical conduct of this research, particularly concerning data handling, has been a paramount consideration.

While the primary datasets utilized in this study are provided in an anonymized or de-identified form by their creators, adherence to their usage agreements and ethical best practices remains crucial. This includes acknowledging the data sources appropriately and ensuring that any further processing or analysis conducted within this research respects the original anonymization and privacy intentions.

Even when working with publicly available, anonymized data, responsible data handling is essential. For instance, if derivative datasets or features were created, care was taken to ensure these did not inadvertently lead to potential re-identification, however unlikely. The focus remains on utilizing the data solely for the stated research objectives, thereby protecting individuals from potential discrimination or harm that could arise from misuse of sensitive information, even in its aggregated or transformed state.

Furthermore, while the original datasets are open-source, researchers involved in this project are bound by agreements and ethical guidelines that prohibit the unauthorized sharing of the dataset. These restrictions are in place to maintain research integrity and comply with institutional review board (IRB) requirements and general ethical research practices.

5.4.2 Sustainability

The sustainability of the developed models and methodologies has also been a key consideration, primarily focusing on computational resources.

The models used in this thesis have been designed with an emphasis on computational efficiency, particularly during the model prediction phase. This approach contributes to sustainability by minimizing energy consumption associated with extensive computational tasks. Lower computational demands

also mean that the research and its potential applications can be more readily adopted and maintained in environments with limited computational infrastructure or financial resources, promoting wider applicability and long-term viability.

Chapter 6

Conclusions and Future work

6.1 Conclusions

This study set out to develop and evaluate a computationally efficient and generalizable algorithm for seizure detection from scalp EEG recordings. By leveraging the ROCKET-based feature transformation algorithms, this work successfully demonstrates a robust methodology that balances high predictive performance with computational efficiency and generalizability, a critical requirement for practical clinical applications.

The primary conclusion of this study is that the Detach-MINIROCKET Ensemble model outperforms both the single Detach-MINIROCKET model and the catch22 feature-based benchmark on in-data and cross-dataset evaluations. This superiority was consistently observed across key performance metrics, most notably in the clinically-relevant event-wise F1 score and a lower False Positive Rate (FPR) per 24 hours. The ensemble's ability to generalize was confirmed through cross-dataset evaluation on the Siena dataset, where it maintained a performance advantage and achieved an F1 score competitive with top-ranked algorithms from a recent seizure detection challenge.

A second key contribution is the demonstration of remarkable computational efficiency. The Detach-MINIROCKET model offered prediction speeds over 17 times faster than the catch22 benchmark, and even the more complex ensemble model remained markedly faster. This efficiency underscores the viability of the proposed approach for real-time or near real-time monitoring systems where low latency is crucial. Furthermore, the sensitivity analysis indicated that the models perform robustly even when trained on substantially reduced dataset sizes, highlighting their potential for use in scenarios with limited data availability.

Finally, this study provides valuable insights into model interpretability through channel relevance analysis. By identifying channels that consistently contribute most (e.g., Fp1, C4) and least (e.g., F8) to seizure detection, this study offers a pathway for future model optimization, feature selection, and potentially a better understanding of the electrographic signatures of seizures as captured by this methodology.

Despite these successes, the study acknowledges that the achieved False Positive Rate, while improved by the ensemble, remains a significant challenge for direct clinical deployment and requires further reduction. Nevertheless, this thesis successfully establishes the Detach-ROCEKT-based framework as a powerful, efficient, and highly generalizable approach for automated seizure detection, providing a solid foundation for future research aimed at clinical translation and the development of neurological diagnostic tools.

6.2 Future work

Building on this study's findings and limitations, future work will focus on enhancing the model's practical utility, robustness, and analytical depth to advance its clinical viability.

A key area for future work is to train and evaluate models on a limited subset of EEG channels. Guided by the channel relevance findings in this thesis (e.g., prioritizing channels like Fp1 and C4 over less informative ones like F8), this approach would further reduce computational complexity and enhance the system's practical use for wearable or resource-constrained EEG devices where a full 19-channel setup is impractical.

While the current ensemble uses a fixed epoch duration, our sensitivity analysis suggested that different temporal windows might capture distinct seizure characteristics. A future iteration of the ensemble could be constructed from base models trained on varying epoch durations, potentially creating a more robust model that is sensitive to patterns across multiple time scales, thereby improving detection accuracy and generalization.

Future research could also refine the feature selection process by exploring adaptive pruning strategies(e.g detach from a feature space that is grown from the primarily detached features). Furthermore, rather than relying on a purely random process, methods to guide the initial kernel generation could be investigated to produce features more tailored to EEG signals and detection tasks.

Finally, to better understand the temporal dynamics of seizures, future work could employ Hidden Markov Models (HMMs) based on detached

features. By using the Detach-Ensemble's probability outputs, an HMM could model the transitions between pre-ictal, ictal, and post-ictal states. This approach could potentially improve temporal coherence, reduce false positives, and enable proactive seizure forecasting by reliably identifying the pre-ictal phase.

References

[1] 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. [Pages 1 and 5.]

[2] R. S. Fisher, J. H. Cross, J. A. French, N. Higurashi, E. Hirsch, F. E. Jansen, L. Lagae, S. L. Moshé, J. Peltola, E. Roulet Perez *et al.*, "Operational classification of seizure types by the international league against epilepsy: position paper of the ilae commission for classification and terminology," *Epilepsia*, vol. 58, no. 4, pp. 522–530, 2017. [Pages 1 and 5.]

[3] World Health Organization, "Epilepsy," February 2023, accessed on [Insert Date of Access Here]. [Pages 1 and 5.]

[4] R. D. Thijs, R. Surges, T. J. O'Brien, and J. W. Sander, *Epilepsy in adults*. Elsevier, 2019, vol. 393, no. 10172. [Pages 1 and 5.]

[5] B. C. Jobst and C. A. Schevon, "Cognitive dysfunction in adults with epilepsy," *CNS spectrums*, vol. 15, no. S8, pp. 12–19, 2010. [Pages 1 and 5.]

[6] K. M. Fiest, J. Dykeman, S. B. Patten, S. Wiebe, G. G. Kaplan, C. J. Maxwell, A. G. M. Bulloch, and N. Jetté, "Depression in epilepsy: a systematic review and meta-analysis," *Neurology*, vol. 80, no. 6, pp. 590–599, 2013. [Pages 1 and 5.]

[7] O. Devinsky, A. Vezzani, T. J. O'Brien, M. Dichter, E. Perucca, W. Löscher, and I. E. Scheffer, "Epilepsy," *Nature reviews Disease primers*, vol. 4, no. 1, pp. 1–23, 2018. [Pages 1 and 6.]

[8] E. Niedermeyer and F. L. Da Silva, *Electroencephalography: Basic principles, clinical applications, and related fields*, 5th ed. Lippincott Williams & Wilkins, 2005. [Pages 1 and 6.]

[9] S. Roy, I. Kiral-Kornek, and S. Harrer, "A review on epileptic seizure detection using machine learning classifiers," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 249–264, 2021. [Pages 1 and 10.]

[10] U. R. Acharya, S. V. Sree, G. Swapna, R. J. Martis, and J. S. Suri, "Automated diagnosis of epileptic eeg using entropies," *Entropy*, vol. 15, no. 10, pp. 4116–4130, 2013. [Pages 2 and 10.]

[11] A. Subasi and M. I. Gursoy, "Practical guide for implementing machine learning for eeg signal processing," *IEEE Sensors Journal*, vol. 19, no. 21, pp. 9649–9667, 2019. [Pages 2 and 10.]

[12] A. Shoeibi, M. Khodatars, N. Ghassemi, M. Jafari, P. Moridian, R. Alizadehsani, M. Panahiazar, F. Khozeimeh, A. Zare, H. Hosseini-Nejad *et al.*, "Epileptic seizures detection using deep learning techniques: A review," *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, p. 5780, 2021. [Pages 2 and 10.]

[13] S. Anwar, A. Al-Jumaily, and K. Khurshid, "Deep learning methods for electroencephalogram (eeg) data analysis: a review," *Multimedia Tools and Applications*, vol. 81, no. 1, pp. 1–34, 2022. [Pages 2 and 10.]

[14] A. Dempster, F. Petitjean, and G. I. Webb, "Rocket: exceptionally fast and accurate time series classification using random convolutional kernels," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, p. 1454–1495, Jul. 2020. doi: 10.1007/s10618-020-00701-z. [Online]. Available: http://dx.doi.org/10.1007/s10618-020-00701-z [Pages 2, 9, 12, 25, and 34.]

[15] A. Dempster, D. F. Schmidt, and G. I. Webb, "Minirocket: A very fast (almost) deterministic transform for time series classification," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining*, ser. KDD '21. ACM, Aug. 2021. doi: 10.1145/3447548.3467231 p. 248–257. [Online]. Available: http://dx.doi.org/10.1145/3447548.3467231 [Pages 2, 9, 13, 25, and 34.]

[16] G. Uribarri, F. Barone, A. Ansuini, and E. Fransén, "Detach-rocket: Sequential feature selection for time series classification with random convolutional kernels," 2024. [Online]. Available: https://arxiv.org/abs/2309.14518 [Pages 2, 9, 13, and 14.]

[17] A. T. Berg, S. F. Berkovic, M. J. Brodie, J. Buchhalter, J. H. Cross, W. van Emde Boas, J. Engel, J. French, T. A. Glauser, G. W. Mathern *et al.*, "Revised terminology and concepts for organization of seizures and epilepsies: report of the ilae commission on classification and terminology, 2005–2009," *Epilepsia*, vol. 51, no. 4, pp. 676–685, 2010. [Page 5.]

[18] H. M. de Boer, S. L. Moshé, E. Gaily, R. Grünewald, W. Mühlnickel, M. Siniatchkin, J. W. Sander, and H. Meinardi, "The global burden and stigma of epilepsy," *Epilepsia*, vol. 49, pp. 50–54, 2008. [Page 5.]

[19] P. E. Smith, "The initial diagnosis of epilepsy," *British journal of hospital medicine*, vol. 76, no. 12, pp. 690–695, 2015. [Page 6.]

[20] W. C. LaFrance Jr and O. Devinsky, "Psychogenic nonepileptic seizures," *Current neurology and neuroscience reports*, vol. 8, no. 4, pp. 306–313, 2008. [Page 6.]

[21] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. Hudspeth, *Principles of neural science*. McGraw-Hill, New York, 2013, vol. 5. [Pages 6 and 7.]

[22] S. Sanei and J. A. Chambers, "Eeg signal processing," 2007. [Page 6.]

[23] S. Noachtar and A. S. Peters, "The semiology of epileptic seizures: a critical review," *Epilepsy Behavior*, vol. 15, no. 1, pp. 2–10, 2009. [Page 6.]

[24] M. Teplan, "Fundamentals of eeg measurement," *Measurement science review*, vol. 2, no. 2, pp. 1–11, 2002. [Pages 6 and 7.]

[25] G. H. Klem, H. O. Lüders, H. H. Jasper, and C. Elger, "The ten-twenty electrode system of the international federation," *Electroencephalography and clinical neurophysiology. Supplement*, vol. 52, pp. 3–6, 1999. [Pages 6 and 7.]

[26] R. Oostenveld and P. Praamstra, "Guidelines for the use of eeg and meg in cognitive neuroscience research: a checklist of good practice," *Clinical Neurophysiology*, vol. 112, no. 4, pp. 575–577, 2001. [Page 6.]

[27] E. K. St Louis and N. Foldvary-Schaefer, "Epilepsy and arousals from sleep," *Sleep medicine reviews*, vol. 13, no. 4, pp. 297–307, 2009. [Page 7.]

[28] M. K. Islam, A. Rastegarnia, and Z. Yang, "Eeg artifacts handling in brain-computer interface applications: a comprehensive review," *Neural Processing Letters*, vol. 48, pp. 877–900, 2018. [Page 7.]

[29] X. Jiang, G.-B. Bian, and Z.-A. Tian, "Removal of artifacts from eeg signals: a review," *Sensors*, vol. 19, no. 5, p. 987, 2019. [Page 7.]

[30] M. Fatourechi, A. Bashashati, R. K. Ward, and G. E. Birch, "Emg and eog artifacts in practical bci systems: a survey," *Clinical neurophysiology*, vol. 118, no. 3, pp. 486–504, 2007. [Page 7.]

[31] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances," *Data Mining and Knowledge Discovery*, vol. 31, pp. 606–660, 2017. [Pages 7, 8, and 23.]

[32] Z. Xing, J. Pei, and E. Keogh, "A brief review on time series analysis," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 40–48, 2010. [Page 7.]

[33] P. Esling and C. Agon, "Time-series data mining," *ACM computing surveys (CSUR)*, vol. 45, no. 1, pp. 1–34, 2012. [Page 8.]

[34] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 7, no. 3, pp. 1–45, 2013. [Page 8.]

[35] L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," pp. 947–956, 2009. [Page 8.]

[36] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data mining and knowledge discovery*, vol. 33, pp. 917–963, 2019. [Page 8.]

[37] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," pp. 1578–1585, 2017. [Page 8.]

[38] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and J. Lines, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, pp. 1936–1962, 2020. [Page 8.]

[39] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: A review," *Data Mining and Knowledge Discovery*, vol. 33, no. 4, pp. 917–963, 2019. [Page 8.]

[40] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," pp. 2114–2124, 2021. [Page 8.]

[41] A. Bagnall, J. Lines, J. Hills, and A. Bostrom, "Time-series classification with COTE: The collective of transformation-based ensembles," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 9, pp. 2522–2535, 2015. [Page 8.]

[42] J. Lines, S. Taylor, and A. Bagnall, "Time series classification with HIVE-COTE: The hierarchical vote collective of transformation-based ensembles," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 5, pp. 1–35, 2018. [Page 8.]

[43] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, and A. Bagnall, "HIVE-COTE 2.0: a new meta ensemble for time series classification," *Machine Learning*, vol. 110, no. 11-12, pp. 3211–3243, 2021. [Page 8.]

[44] B. D. Fulcher, M. A. Little, and N. S. Jones, "Highly comparative time-series analysis: the empirical structure of time series and their methods," *Journal of the Royal Society Interface*, vol. 10, no. 83, p. 20130048, 2013. [Pages 8, 9, and 23.]

[45] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "catch22: Canonical time-series characteristics," *Data Mining and Knowledge Discovery*, vol. 33, pp. 1821–1852, 2019. [Pages 8, 9, 15, and 23.]

[46] A. P. Ruiz, M. Flynn, J. Large, M. Middlehurst, and A. Bagnall, "The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances," *Data Mining and Knowledge Discovery*, vol. 35, no. 2, pp. 401–449, 2021. [Page 9.]

[47] C. W. Tan, A. Dempster, C. Bergmeir, and G. I. Webb, "Multirocket: multiple pooled representations from random convolutional kernels for fast and accurate time series classification," *Data Mining and Knowledge Discovery*, vol. 36, no. 5, pp. 1738–1762, 2022. [Page 9.]

[48] A. Solana, E. Fransén, and G. Uribarri, "Classification of raw meg/eeg data with detach-rocket ensemble: An improved rocket algorithm for multivariate time series analysis," 2024. [Online]. Available: https://arxiv.org/abs/2408.02760 [Pages 9, 14, 18, and 29.]

[49] S. Usmankhujaev, C.-H. Lee, and L. CKM, "A review of eeg-based automatic seizure detection methods," *Applied Sciences*, vol. 11, no. 18, p. 8409, 2021. [Page 10.]

[50] R. Hussein, N. R. Pal, M. Abo-Zahhad, and A. Shalaby, "Epileptic seizures detection in eeg signals using a hybrid of comprehensive feature extraction and machine learning techniques," *Biomedical Signal Processing and Control*, vol. 59, p. 101910, 2020. [Page 10.]

[51] A. Gupta, P. Singh, and M. Karlekar, "A review on eeg based epileptic seizure detection using signal processing and machine learning techniques," *Intelligent Systems with Applications*, vol. 14, p. 200064, 2021. [Page 10.]

[52] I. Kiral-Kornek, S. Roy, E. Nurse, and S. Harrer, "Machine learning-based epileptic seizure detection methods using wavelet and emd-based decomposition techniques: A review," *Sensors*, vol. 23, no. 1, p. 484, 2023. [Page 10.]

[53] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "Attention-based deep models for sleep stage classification with eog and eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1119–1130, 2021. [Page 10.]

[54] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "A deep convolutional neural network for automated detection of epileptic seizures using eeg signals," *Computers in biology and medicine*, vol. 100, pp. 270–278, 2018. [Page 10.]

[55] M. Zhou, C. Tian, R. Cao, B. Wang, Y. Niu, T. Hu, X. Guo, and J. Xiang, "Patient-specific seizure detection in long-term eeg using a convolutional neural network with transfer learning from a common

model," *Biomedical Signal Processing and Control*, vol. 46, pp. 104–111, 2018. [Page 10.]

[56] M. Golmohammadi, S. S. Ziyabari, V. Shah, E. von Weltin, I. Obeid, and J. Picone, "A review of automated eeg analysis for seizure detection," *Journal of clinical neurophysiology*, vol. 36, no. 4, pp. 249–261, 2019. [Page 10.]

[57] R. Hussein, N. R. Pal, R. K. Ward, and Z. J. Wang, "Human interpretable deep learning for epileptic seizure detection," *IEEE Access*, vol. 7, pp. 78 122–78 132, 2019. [Page 10.]

[58] 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. 22–37, 2018. [Page 10.]

[59] M. S. Aslam, M. U. G. Khan, M. N. Asghar, and S. Lee, "A novel cnn-lstm based framework for epileptic seizure prediction," *Artificial Intelligence in Medicine*, vol. 123, p. 102231, 2022. [Page 10.]

[60] K. Indurani and P. Vandana, "Time-attention based convolutional and recurrent neural network for epileptic seizure prediction using eeg signals," *Biomedical Signal Processing and Control*, vol. 80, p. 104339, 2023. [Page 10.]

[61] W. Zhao, A. F. Struck, S. D. Lhatoo, and G.-Q. Zhang, "Automated seizure detection using transformer models on multi-channel eegs," *Computers in Biology and Medicine*, vol. 153, p. 106497, 2023. [Page 10.]

[62] Y. Tian, Z. Jia, and Y. Li, "A hybrid model combining cnn and transformer for seizure detection based on brain connectivity," *Biomedical Signal Processing and Control*, vol. 69, p. 102857, 2021. [Page 10.]

[63] Y. Wang, C. Zou, and J. Liu, "Transformer-based seizure prediction system using dynamic multi-graph convolutional network," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1224–1234, 2022. [Page 10.]

[64] A. S. Koshiyama and Fsub, "Unsupervised multivariate time-series transformers for seizure identification on eeg," *arXiv preprint arXiv:2301.03470*, 2023. [Page 10.]

[65] M. Gerster, A. I. Humayun, Z. Yang, N. Liu, H. Wu, F. Cong, O. Grinenko, and C. Lyu, "Tsd: Transformers for seizure detection," *arXiv preprint arXiv:2301.10138*, 2023. [Page 10.]

[66] Y. Liu, Y. Wang, and J. Wang, "Automatic seizure detection based on stockwell transform and transformer," *Biomedical Signal Processing and Control*, vol. 87, p. 105377, 2024. [Page 10.]

[67] Y. Zhang and H. Liu, "Efficient eeg feature learning model combining random convolutional kernel with wavelet scattering for seizure detection," *International Journal of Neural Systems*, 2024. [Page 10.]

[68] J. Lundy and J. M. O'Toole, "Automated neonatal eeg classification using random convolutional kernel transform (rocket)," *IRISH SIGNALS AND SYSTEMS CONFERENCE (ISSC)*, pp. 1–6, 2021. [Page 10.]

[69] K. Lee, H. Jeong, S. Kim, D. Yang, H.-C. Kang, and E. Choi, "Real-time seizure detection using eeg: A comprehensive comparison of recent approaches under a realistic setting," *Proceedings of Machine Learning Research*, vol. 174, pp. 311–337, 2022. [Page 10.]

[70] G. Saleh, K. N. Fountas, and E. Z. Kapsalaki, "High accuracy of epileptic seizure detection using tiny machine learning technology for implantable closed-loop neurostimulation systems," *Bioengineering*, vol. 10, no. 1, p. 103, 2023. [Page 10.]

[71] V. Shah, E. von Weltin, S. Lopez, J. McHugh, L. Veloso, M. Golmohammadi, I. Obeid, and J. Picone, "The temple university hospital seizure detection corpus," *Frontiers in Neuroinformatics*, vol. 12, p. 83, 2018. doi: 10.3389/fninf.2018.00083. [Online]. Available: https://doi.org/10.3389/fninf.2018.00083 [Pages ix, 11, and 12.]

[72] I. Obeid and J. Picone, "The temple university hospital eeg data corpus," in *Augmentation of Brain Function: Facts, Fiction and Controversy. Volume I: Brain-Machine Interfaces*, 1st ed., C. Guger, B. Z. Allison, and G. Edlinger, Eds. Lausanne, Switzerland: Frontiers Media S.A., 2018, pp. 394–398. [Page 11.]

[73] P. Detti, "Siena scalp eeg database (version 1.0.0)," https://doi.org/10.13026/5d4a-j060, 2020, physioNet. [Page 11.]

[74] P. Detti, G. Vatti, and G. Z. M. de Lara, "EEG Synchronization Analysis for Seizure Prediction: A Study on Data of Noninvasive Recordings," *Processes*, vol. 8, no. 7, p. 846, 2020. doi: 10.3390/pr8070846. [Online]. Available: https://doi.org/10.3390/pr8070846 [Page 11.]

[75] J. Dan, U. Pale, A. Amirshahi, W. Cappelletti, T. M. Ingolfsson, X. Wang, A. Cossettini, A. Bernini, L. Benini, S. Beniczky, D. Atienza, and P. Ryvlin, "Szcore: Seizure community open-source research evaluation framework for the validation of electroencephalography-based automated seizure detection algorithms," *Epilepsia*, vol. n/a, no. n/a. doi: https://doi.org/10.1111/epi.18113 [Page 12.]

[76] J. Dan *et al.*, "Library for measuring performance of time series classification," https://github.com/esl-epfl/timescoring, 2023, accessed: 2025-05-04. [Page 15.]

[77] R. D. Pascual-Marqui, D. Lehmann, T. Koenig, K. Kochi, M. C. Merlo, D. Hell, and M. Koukkou, "Low resolution brain electromagnetic tomography (loreta) functional imaging in acute, neuroleptic-naive, first-episode, productive schizophrenia," *Psychiatry Research: Neuroimaging*, vol. 90, no. 3, pp. 169–179, 1999. doi: https://doi.org/10.1016/S0925-4927(99)00013-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092549279900013X [Page 33.]

[78] Y. Lai, W. van Drongelen, L. Ding, K. E. Hecox, V. L. Towle, D. M. Frim, and B. He, "Estimation of in vivo human brain-to-skull conductivity ratio from simultaneous extra- and intra-cranial electrical potential recordings," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 116, no. 2, pp. 456–465, Feb 2005. doi: 10.1016/j.clinph.2004.08.017. [Online]. Available: https://doi.org/10.1016/j.clinph.2004.08.017 [Page 33.]

[79] T. M. Ingolfsson, S. Benatti, X. Wang, A. Bernini, P. Ducouret, P. Ryvlin, S. Beniczky, L. Benini, and A. Cossettini, "Minimizing artifact-induced false-alarms for seizure detection in wearable eeg devices with gradient-boosted tree classifiers," *Scientific Reports*, vol. 14, no. 1, p. 2980, 2024. [Page 35.]

TRITA – EECS-EX 2024:0000
Stockholm, Sweden 2025

www.kth.se