UNIVERSIDADE D
COIMBRA

Joana Carolina Flórido Batista

On the clinical acceptance of EEG seizure prediction methodologies

Thesis submitted to the Faculty of Science and Technology of the University of Coimbra for the degree of Master in Biomedical Engineering with specialization in Clinical Informatics and Bioinformatic, supervised by Prof. Dr. César Teixeira and MSc Mauro Pinto.

September 2022

1 2  9 0

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Joana Carolina Flórido Batista

On the clinical acceptance of EEG seizure prediction methodologies

Thesis submitted to the
University of Coimbra for the degree of
Master in Biomedical Engineering

Supervisors:
Prof. Dr. César Alexandre Domingues Teixeira (CISUC)
MSc Mauro Filipe da Silva Pinto (CISUC)

Coimbra, 2022

This work was developped in collaboration with:

CISUC - Center for Informatics and Systems of the University of Coimbra

This work was supported by the Portuguese Foundation for Science and Technology through projects CISUC (UID/CEC/00326/2020) and RECoD (PTDC/EEI-EEE/5788/2020).

Agradecimentos

Em primeiro lugar, gostaria de agradecer aos meus orientadores Professor Doutor César Teixeira e Mestre Mauro Pinto, por toda a dedicação, disponibilidade, partilha de conhecimentos e apoio prestado ao longo da realização deste projeto.

Gostaria também de agradecer aos meus colegas de laboratório por toda a ajuda e partilha de ideias.

Deixo o meu maior agradecimento à minha família, por todo o apoio. Em especial ao meus pais e à minha mana, por tornarem estes 5 anos possíveis. Pelas noites em branco à espera da prometida mensagem de que tinha chegado bem a casa. Pela paciência, apoio incondicional e motivação até nas horas de maior desespero. Obrigada!

Por fim, a todos aqueles que me acompanharam nesta aventura, aos amigos que levarei comigo para a vida. Obrigada por tornarem Coimbra ainda mais especial. Foram 5 anos de desafios, 5 anos de amizades, 5 anos de memórias ... 5 anos de Coimbra!

"Everything is hard before it is easy."

GOETHE

Resumo

Sensivelmente um terço dos doentes epiléticos são incapazes de atingir o controlo das crises através da administração de medicamentos antiepiléticos. Em situações em que as crises epiléticas são impossíveis de controlar, a previsão de crises desempenha um papel fundamental no planeamento clínico e terapêutico, fornecendo novas opções de tratamentos, como dispositivos de alerta ou de intervenção. Estes sistemas têm a potencialidade de melhorar a qualidade de vida dos doentes suscetíveis à ocorrência súbita de crises.

No entanto, a falta de interpretabilidade e explicabilidade das abordagens utilizadas nesta área constitui um obstáculo na aplicação clínica das metodologias de previsão e sistemas de intervenção desenvolvidos. Os métodos atualmente utilizados na literatura são maioritariamente baseados em sistemas complexos, difíceis de serem interpretados e de garantir a confiança dos clínicos.

O presente trabalho teve como objetivo o desenvolvimento de metodologias capazes de prever crises epiléticas ao mesmo tempo que garantem a confiança dos clínicos, cientistas de dados e doentes.

Foi desenvolvido um algoritmo para a previsão de crises epiléticas, considerando 40 doentes da base de dados EPILEPSIAE, resistentes à medicação antiepilética. Através da metodologia proposta, foram obtidos resultados de $0{,}34 \pm 0{,}35$ para a sensibilidade e de $1{,}78 \pm 1{,}95$ para o FPR/h. Sendo que 40% dos modelos desenvolvidos apresentaram uma performance estatisticamente significativa.

Posteriormente, diferentes estratégias de explicabilidade foram aplicadas de forma a aumentar a confiança nas decisões do modelo. As explicações elaboradas basearam-se nas cinco lições extraídas de um trabalho previamente desenvolvido neste laboratório. As curvas de regularização ao longo do tempo foram analisadas para todos os doentes, e comportamentos típicos do modelo foram observados um número de vezes estatisticamente significativo. Foi ainda avaliado o impacto de cada uma das bandas espectrais sobre a capacidade de previsão do modelo, o que permitiu concluir

que, em cenários específicos, diferentes conjuntos de características podem provocar comportamentos completamente distintos no classificador.

Com este estudo, foi possível concluir que para algoritmos de previsão de crises epiléticas, a explicabilidade não deve simplesmente explicar as decisões do classificador, é também necessário melhorar os modelos desenvolvidos, rever os pressupostos e elaborar uma formulação mais completa do problema, de forma a garantir uma maior confiança sobre as metodologias desenvolvidas.

Palavras-chave: Epilepsia, Machine learning, Previsão de crises, Interpretabilidade, Explicabilidade

Abstract

Almost one-third of epileptic patients fail to achieve seizure control through antiepileptic drug administration. In the scarcity of completely controlling a patient's epilepsy, seizure prediction plays a significant role in clinical management and treatment, providing new therapeutic options such as warning or intervention devices. These systems would attempt to improve the quality of life of patients who are susceptible to the sudden occurrence of seizures.

However, the lack of interpretability and explainability of the seizure prediction approaches constitutes an obstacle to the clinical applicability of the proposed prediction methodologies and intervention devices. The current state-of-the-art methods are mainly based on complex models that are difficult to trust by domain experts.

The present work aimed to explore methodologies capable of predicting epileptic seizures in ways that guarantee trust to data scientists, clinicians, and patients.

Considering 40 drug-resistant epilepsy patients from the EPILEPSIAE database, a patient-specific seizure prediction algorithm was developed. The proposed methodology achieved 0.34 ± 0.35 for sensitivity and 1.78 ± 1.95 for FPR/h, where 40% of patient models performed above chance.

Afterwards, different explaining strategies were employed to increase trust in the models' decisions. These explanations were based on five lessons extracted from a prior work developed by the local research team. The patients' time plots were inspected, and typical model behaviors were found in a statistically significant number. The impact of each classical Electroencephalogram (EEG) spectral band over the model prediction was also evaluated, and it was concluded that, in specific scenarios, different sets of features might produce an entirely distinct behavior in the classifiers' output.

With this study, it was possible to conclude that for seizure prediction algorithms, explainability should not simply explain the model's decision. It is necessary to

improve the developed models, review used assumptions, and create a completer problem formulation to gain trust.

Keywords: Epilepsy, Machine learning, Seizure prediction, Interpretability, Explainability

Contents

List of Figures **xviii**

List of Tables **xxii**

List of Abbreviations **xxiv**

1 Introduction **1**

1.1 Motivation 1

1.2 Context 1

1.2.1 General goal 1

1.2.2 Seizure prediction limitations 2

1.2.3 The importance of explainability 3

1.3 Objectives 3

1.4 Structure 4

2 Background Concepts **5**

2.1 Epilepsy 5

2.1.1 Definition 5

2.1.2 Classification 6

2.1.2.1 Seizure Type 7

2.1.2.2 Epilepsy Type 8

2.1.2.3 Epilepsy Syndrome 9

2.1.3 Seizure clusters 10

2.1.4 Treatment 10

2.1.4.1 Drug-Resistant Epilepsy (DRE) 11

2.2 Electroencephalogram (EEG) 11

2.2.1 Overview 11

2.2.2 Signal acquisition 13

2.2.3 EEG seizure period division 16

2.3 Seizure Prediction . . . 16
2.3.1 Seizure prediction characterization . . . 19
2.3.2 Performance evaluation . . . 20
2.3.3 Statistical Validation . . . 23
2.3.3.1 Analytical random predictor . . . 23
2.3.3.2 Surrogate seizure predictor . . . 24
2.3.4 Concept drift and class imbalance . . . 26
2.4 Explainability . . . 26
2.4.1 Taxonomy . . . 30
2.4.2 Evaluation . . . 30
2.4.3 Explainability methods . . . 32
2.4.3.1 Global model-agnostic methods . . . 33
2.4.3.2 Local model-agnostic methods . . . 36
2.4.3.3 Example-based methods . . . 39
2.4.3.4 Deep Learning interpretation . . . 40
2.4.4 Grounded Theory (GT) . . . 41
2.5 Summary . . . 43

3 State of the art **45**
3.1 Seizure Prediction . . . 45
3.1.1 Framework overview . . . 45
3.1.2 Signal acquisition . . . 47
3.1.3 Signal pre-processing . . . 50
3.1.4 Feature Extraction . . . 52
3.1.5 Feature Selection . . . 55
3.1.6 Classification . . . 55
3.1.7 Regularization . . . 58
3.1.8 Performance evaluation . . . 59
3.2 Explainability . . . 59
3.2.1 Intrinsically interpretable models . . . 59
3.2.2 Non-interpretable models . . . 60
3.2.2.1 Explainability in Epilepsy studies . . . 60
3.2.2.2 Explainability in other EEG studies . . . 64
3.3 Summary . . . 65

4 Methodology **69**
4.1 Pipeline Overview . . . 69
4.2 Seizure Prediction . . . 70

4.2.1 Data . . . 70
4.2.2 Pre-processing . . . 71
4.2.3 Feature Extraction . . . 72
4.2.4 Data Splitting . . . 73
4.2.5 Training . . . 74
4.2.5.1 Class labeling . . . 74
4.2.5.2 Class balancing . . . 74
4.2.5.3 Feature Standardization . . . 75
4.2.5.4 Feature Selection . . . 75
4.2.5.5 Classifier . . . 75
4.2.5.6 Grid-Search . . . 76
4.2.6 Testing . . . 77
4.2.7 Post-processing . . . 77
4.2.8 Performance Evaluation . . . 79
4.3 Explainability . . . 79
4.3.1 Prior work . . . 79
4.3.2 Adopted Methodology . . . 82
4.3.2.1 Analysis and hypotheses formulation . . . 82
4.3.2.2 Statistical validation . . . 83

5 Results and Discussion **87**
5.1 Seizure Prediction . . . 87
5.1.1 Training phase . . . 87
5.1.2 Testing phase . . . 89
5.1.3 Comparative analysis with other studies . . . 91
5.2 Explainability . . . 93
5.2.1 Time analysis . . . 93
5.2.1.1 Statistical validation . . . 101
5.2.2 Feature analysis . . . 103

6 Conclusion **113**

Bibliography **115**

Appendices **129**
A Features Description . . . 131

List of Figures

2.1 ILAE 2017 framework for classification of epilepsies. *Denotes onset of seizure. Adapted from: Sheffer et al. 2017 [1]. 7
2.2 The basic ILAE 2017 operational classification of seizure types. [1]Definitions, other seizure types and descriptors are listed in the accompanying paper and glossary of terms. [2] Due to inadequate information or inability to place in other categories. Adapted from: Sheffer et al. 2017 [2]. 7
2.3 Categorization of EEG activity. 13
2.4 International 10-20 system for placement of scalp EEG electrodes. In (a) is shown the standard positions and names of the electrodes. In (b) is represented a bipolar montage and in (c) a referential montage. Adapted from: Varsavsky et al. 2011 [3]. 15
2.5 Different types of invasive electrodes. Extracted from: [4]. 16
2.6 Different periods of an epileptic seizure annotated on the EEG signal. All four states of ictal, preictal, ictal, postictal and interictal are colour coded. Source: Moghim et al. [5]. 17
2.7 Visual representation of SPH and SOP. Adapted from: Winterhalder et al. [6]. 19
2.8 Visual representation of true and false alarms in seizure prediction, considering SPH and SOP. 21
2.9 Representation of the original seizure times and the surrogate times bootstrapped from the inter-seizure intervals. The arbitrary onset times for the surrogates are originated through a uniform distribution and are represented by the dashed vertical lines. Source: Schelter et al. 2008 [7]. 25
2.10 The trade-off between interpretability and accuracy of some relevant Machine Learning (ML) models. 27
2.11 The big picture of explainable machine learning. 28

2.12 Taxonomy mind-map of Machine Learning Interpretability Techniques. Adapted from: Linardatos et al. [8]. 31
2.13 Taxonomy of evaluation approaches for interpretability. Adapted from: Doshi-Velez et al. [9]. 32
2.14 Partial Dependence Plot (PDP) for the prediction count model of bicycle renting of weather features (temperature, humidity and wind speed). Source: Molnar et al. [10]. 33
2.15 Prototypes and criticisms for a data distribution with two features. Source: Molnar et al. [10]. 36
2.16 Individual Conditional Expectation (ICE) plots for the prediction count model of bicycle renting of weather features (temperature, humidity and wind speed). Source: Molnar et al. [10]. 37
2.17 Shapley values regarding an instance from the prediction count model of bicycle daily renting. Source: Molnar et al. [10]. 39
2.18 Representation of a linear model with one feature. Trained once with the full data and once without the influential feature. Source: Molnar et al. [10]. 40
2.19 Grounded Theory (GT) flow chart. 43

3.1 General framework of seizure prediction studies. Adapted from: Rasheed et al. [11]. 46
3.2 EEG features most commonly used in seizure prediction studies, categorized in terms of number of channels and linearity. 53

4.1 General overview of the proposed pipeline. 69
4.2 List of the linear univariate features extracted from the EEG in this work. A total of 59 features in the time and frequency domains were computed from each window on 19 EEG electrodes. 73
4.3 Procedure applied to prepare data for train and test. 74
4.4 Random undersampling of interictal class respecting the sequential chronology of samples. Red colored samples correspond to interictal samples randomly chosen from each group. Only one hypothetical seizure with 10 preictal samples is illustrated. 75
4.5 Grid-search procedure implemented to select the optimal training parameters for each preictal period. 77
4.6 Procedure applied to train an test the seizure prediction model. . . . 78

4.7 Visual representation of the firing power technique implemented. Given a certain threshold (dashed line), an alarm is only triggered when the firing power exceeds its value and is at least one refractory period separated from the last generated one. Two false alarms and one true alarm are illustrated. 78
4.8 Representation of the distinction between the statistical evaluation with one and two binomial distributions. 84
4.9 Procedure applied for statistical validation when two binomial distributions are used. 85
4.10 Procedure applied for statistical validation when one binomial distribution is used. 86

5.1 Plot of the ensemble of SVMs' decisions over time for patient 8902 seizure #4. Each SVM decision is in grey and the ensemble voting system is in black. The vigilance state is also represented. 94
5.2 Plot of the ensemble of SVMs' decisions over time for patient 93902 seizure #6. Each SVM decision is in grey, and the ensemble voting system is in black. The rectangle drawn in the figure represents the refractory time to illustrate why no true alarm was raised during the preictal period. The vigilance state is also represented. 95
5.3 **(a)** Plot of the ensemble of SVMs' decisions over time for patient 114902 seizure #7. Each SVM decision is in grey, and the ensemble voting system is in black. The vigilance state is also represented. **(c)** Plot of the model decisions over time for patient 114902 considering all SOPs. **(c)** Testing performance (SS, FPR/h) for patient 114902 considering each SOP. 96
5.4 Visual representation of the circadian forecasting model implementation. One testing seizure is predicted if its onset occurs during the seizure risk interval. 97
5.5 Plot of the ensemble of SVMs' decisions over time for patient 94402 seizure #6. Each SVM decision is in grey and the ensemble voting system is in black. The vigilance state is also represented. 99
5.6 Plot of the model decisions over time for patient 8902 considering all seizure. Only the ensemble voting system is depicted in black. The dashed orange line represents the preictal period of each seizure.The vigilance state is also represented. 100

5.7 Plot of the ensemble of SVMs' decisions over time for patient 98202 seizure #4. Each SVM decision is in grey and the ensemble voting system is in black. The vigilance state is also represented. 101
5.8 Relative frequency of the selected features and channels for patient 8902. 104
5.9 Beeswarm summary plot of Shap Values for patient 8902. 104
5.10 Firing power time plots of patient 8902 for different models considering distinct sets of features. The black line represents the original model, the remaining concern SVM models trained with only a determined spectral band or without gamma band. All the testing seizures are represented (**(a)**, and **(b)**). 106
5.11 Firing power time plots of patient 94402 for different models considering distinct sets of features: all-feature model (left column), gamma-related model (middle column), and without gamma features model (right column). All the testing seizures are represented. 107
5.12 Firing power time plots of patient 94402 for different models considering distinct sets of fetaures. The black line represents the original model, the remaining concern SVM models trained with only a determined spectral band or without gamma band. All the testing seizures are represented (**(a)**, **(b)**, **(c)** , and **(d)**). 109
5.13 Firing power time plots of patient 98202 for different models considering distinct sets of features. The black line represents the original model, the remaining concern SVM models trained with only a determined spectral band or without gamma band. All the testing seizures are represented (**(a)**, **(b)**, **(c)** , **(d)**, and **(e)**). 110
5.14 Firing power time plots of patient 112802 for different models considering distinct sets of fetaures. The black line represents the original model, the remaining concern SVM models trained with only a determined spectral band or without gamma band. All the seizures are represented. 111

List of Tables

2.1 Confusion matrix for evaluation of sample performance in Machine Learning problems. 20

3.1 Overview of the signal acquisition aspects underlying EEG seizure prediction studies over the past 10 years. 48
3.2 Overview of the pre-processing procedures adopted in the EEG seizure prediction studies over the past 10 years. 51
3.3 Overview of the features type used in EEG seizure prediction studies over the past ten years. 54
3.4 Overview of the classification, regularization and performance evaluation characteristics used in EEG seizure prediction studies over the past ten years. 56
3.5 Studies associated with explainability in Epilepsy and EEG decoding and classification. 66

4.1 Information for the 40 studied patients. 70
4.1 Information for the 40 studied patients. 71

5.1 Training parameters and performance obtained for each patient. . . . 88
5.2 Testing performance obtained for each patient. 90
5.3 Seizure prediction performance for studies under comparison. 91
5.4 Sensitivity and time under warning for circadian forecasting model, and respective comparison with the approach under study for each patient. 98
5.5 Analysis of the presence of the identified patterns in each patient. . . 102
5.6 Parameters used in the statistical validation, the total number of occurrences of each phenomenon, and the final result (statistically valid or not). 103

5.7 Testing performance obtained for each patient considering all approaches. 112

List of Abbreviations

AED Anti-Epileptic Drug. 1, 10, 11, 22, 26
ALE Accumulated Local Effects. 34
ANFIS Adaptive Neuro-Fuzzy Inference Systems. 56
ANN Artificial Neural Network. 56
AUC Area Under the Curve. 59

BLDA Bayesian Linear Discriminant Analysis. 56

CNN Convolutional Neural Network. 55, 57, 58, 64, 80
CSP Common Spatial Pattern. 51, 54

DL Deep Learning. 46, 47, 55, 60, 61
DRE Drug-Resistant Epilepsy. xiv, 1, 10, 11, 22, 43, 70, 87
DWT Discrete Wavelet Transform. 132

EEG Electroencephalogram. xii, xiv, xv, xxii, 1, 5, 8, 11, 12, 13, 16, 18, 25, 43, 45, 46, 47, 48, 50, 51, 52, 61, 62, 64, 66, 70, 72, 80, 83, 103, 105
EMD Empirical Mode Decomposition. 51
EMG Electromyogram. 64
EOG Electrooculogram. 64

FDA Food and Drug Administration. 74
FFT Fast Fourier Transform. 132
FIR Finite Impulse Response. 50
FOIA Focal Onset Impaired Awareness. 8, 18
FPR/h False Positive Rate per Hour. x, xii, 20, 21, 22, 23, 43, 59, 80, 113

GA Genetic Algorithm. 55
GAN Generative Adversarial Network. 56
GT Grounded Theory. xv, xix, 41, 43, 80

IBE International Bureau for Epilepsy. 5

ICE Individual Conditional Expectation. xix, 36, 37, 80
iEEG intracranial EEG. 13, 14, 15, 43, 48, 105
IIR Infinite Impulse Response. 50
ILAE International League Against Epilepsy. 5, 6, 7, 11
IT Intervention Time. 19

KNN K-Nearest Neighbors. 39, 56

LIME Local Interpretable Model-agnostic Explanations. 37, 80
LOOCV Leave-One-Out Cross-Validation. 76
LSTM Long Short-Term Memory. 58

mDAD maximum Difference Amplitude Distribution of histogram. 55
ML Machine Learning. xviii, 3, 24, 27, 28, 30, 35, 39, 45, 46, 80, 81, 113
MMD Maximum Mean Discrepancy. 35
MMD-AAE Maximum Mean Discrepancy-Adversarial Autoencoders. 56
mRMR minimum Redundance Maximum Relevance. 55

NN Neural Network. 56

PCA Principal Component Analysis. 55
PDP Partial Dependence Plot. xix, 33, 34, 36, 80
PE Permutation Entropy. 62
PSG Polysomnography. 66

RNN Recurrent Neural Network. 58

SEF Spectral Edge Frequency. 132
SEP Spectral Edge Power. 132
SHAP SHapley Additive exPlanations. 61
SOM Self-Organizing Map. 56
SOP Seizure Occurrence Period. 2, 19, 20, 22, 23, 24, 43, 50, 52, 78, 87
SP Specificity. 20, 76
SPH Seizure Prediction Horizon. 2, 19, 22, 23, 43, 44, 50, 52, 78
SS Sensitivity. 20, 23, 76
SVM Support Vector Machines. xx, 55, 57, 66, 75, 76, 77, 95, 96, 97, 103, 113

TLE Temporal Lobe Epilepsy. 9

1

Introduction

1.1 Motivation

Epilepsy is one of the most common neurological diseases, affecting over 50 million people worldwide. This condition is expressed by atypical brain activity that results in seizures or unusual behavior, sensations and sometimes loss of awareness. This abnormal activity leads to different neurological, cognitive, psychological and social consequences [12].

The first-line treatment for epilepsy is Anti-Epileptic Drugs (AEDs). Nevertheless, almost one-third of patients fail to achieve seizure control with medication alone, being considered Drug-Resistant Epilepsy (DRE) patients [13, 14]. These patients are at a higher risk of developing various psychological problems, such as depression, anxiety, psychosis, and, in the worst scenario, premature death [14, 15]. Although epilepsy surgery is a well-established treatment for DRE patients, only a small amount of patients are eligible for this therapy [15].

In the scarcity of completely controlling a patient's epilepsy, seizure prediction plays a significant role in clinical management and treatment. This approach improves the quality of life of patients who are susceptible to the sudden occurrence of seizures.

1.2 Context

1.2.1 General goal

The seizure prediction field aims to develop an algorithm capable of anticipating an epileptic seizure by raising an alarm before the seizure onset. This field has been moving forward, assuming the existence of a preictal period that Electroencephalogram (EEG) signals can capture. The preictal is a transitional period that precedes the seizure, on which the entire seizure prediction area is grounded [6].

The goal is to design a system able to read online data and properly notify the patient regarding a seizure that will arise on a well-defined occurrence period (SOP) with a predefined horizon (SPH), which must allow enough time to take action. An accurate system may provide new therapeutic options such as warning devices that enable the patient to avoid dangerous situations or even intervention devices capable of controlling the seizure by delivering anticonvulsive drugs or triggering electric stimuli [6, 16].

1.2.2 Seizure prediction limitations

Although the preliminary work on the seizure prediction field dates back to the 1970s, with improving advances over the years, current approaches present numerous limitations which should not be neglected.

The EEG is a complex signal not fully understood by the scientific community. Additionally, the EEG databases are mainly collected from patients during pre-surgical monitoring, which does not reflect actual seizure activity. Long-term EEG recordings, comprising several months or years, acquired in an everyday routine, represent a step forward in the clinical viability of the designed methodologies [16–18].

Concerning the preictal period, it is worth noting that it is the most difficult one to determine and manually annotate by experts since it is associated with substantial heterogeneity. Therefore, no standard or optimal value has yet been defined for the duration of the preictal period. Indeed, there is evidence that this period may vary among patients and between seizures from the same patient. Therefore, the complex nature of this state represents a significant challenge for seizure prediction [17–19].

Class imbalance is another critical issue. In the context of seizure prediction, seizures are relatively rare events, leading to a substantially longer interictal period than the preictal. This issue may induce a specialization of the classifier over the interictal class [17].

Concept drifts constitute another challenging problem. They occur as confounding factors in the EEG signal and may adversely impact the performance of the seizure prediction models. The referred concepts comprise alterations in the brain dynamics depending on exogenous and endogenous factors, such as changes in behavior and mood, cognitive disturbances, circadian rhythm (sleep-wake cycle, time of the day, week, month and year), medication, and others [6, 19–21].

Regarding the seizure prediction methodology, despite the existence of a broad pipeline, there is a great variety of approaches due to the application of different

methods and parameters.

The lack of interpretability and explainability of the seizure prediction approaches also constitutes an obstacle to the clinical applicability of the proposed prediction methodologies and intervention devices.

1.2.3 The importance of explainability

Throughout the last decades, the appearance of more complex algorithms and their deployment in sectors such as healthcare have led to the emergence of explainability and interpretability areas. The most recent legislation is also responsible for the increased interest in this field. In 2018, the European Union's General Data Protection Regulation (GDPR) forced the industries to explain any automatic decision-making process [22, 23].

The deployment of ML models in healthcare has increased interest in optimized systems at the performance level and other essential criteria, including safety, trustworthiness, fairness, robustness, and the right to explanations [8, 9].

Regarding the seizure prediction field, few predictive methodologies and intervention devices have been clinically approved. A great skepticism respecting machine learning models may result from the complexity of interpreting models' decisions. Indeed, models used in clinical trials, such as Neurovista advisory study [24] and intervention devices like the RNS system [25], did not apply the most potent state-of-the-art tools. It proves the skepticism regarding the most complex methodologies and the need for human-understandable explanations.

Therefore, although a given methodology eventually makes incorrect decisions (miss a seizure or raise a false alarm), it is still trustful if it is possible to explain its errors. It is believed that an accurate methodology is the one that we trust.

1.3 Objectives

This project aims to explore methodologies capable of predicting epileptic seizures in ways that guarantee trust to data scientists, clinicians, and patients. Towards this purpose, using long-term EEG data and Machine Learning algorithms, the expected contributions of this thesis are the following:

- Development of a patient-specific methodology for seizure prediction using scalp EEG signals from the European Epilepsy Database (EPILEPSIAE).
- Development and evaluation of several explanations to explain and increase trust in the model's prediction decisions.

1.4 Structure

Besides the introduction, the present document contains five more chapters structured as follows.

Chapter 2 introduces background concepts related to Epilepsy, EEG, seizure predictions, and explainability.

Chapter 3 presents state of the art concerning EEG-based seizure prediction and explainability studies.

Chapter 4 describes the followed steps concerning the primary goal of the present work.

Chapter 5 reports the results obtained from the proposed methodology, along with their analysis and discussion.

Chapter 6 presents a conclusion and addresses future work in this field of study.

2

Background Concepts

This chapter introduces the fundamental concepts required to understand this document. Section 2.1 presents some definitions associated with epilepsy and its classification. Section 2.2 introduces some concepts related to Electroencephalogram (EEG) signal and its characterization. Section 2.3 includes some theoretical insights into the seizure prediction field. Section 2.4 presents an overview of explainability, including a brief description of some explainable methods.

2.1 Epilepsy

2.1.1 Definition

Epilepsy is one of the most common neurological diseases, affecting over 50 million people worldwide of all ages and sex. Epilepsy is characterized by atypical brain activity that results in seizures or unusual behavior, sensations and sometimes loss of awareness. This abnormal activity leads to different neurological, cognitive, psychological and social consequences [12].

In 2005, a conceptual definition for epilepsy was formulated by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE) [26]: *"Epilepsy is a disorder of the brain characterized by an enduring predisposition to generate epileptic seizures and by the neurobiological, cognitive, psychological, and social consequences of this condition. The definition of epilepsy requires the occurrence of at least one epileptic seizure."*. In turn, an epileptic seizure is defined as *"a transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain."*

Later, in 2014, the ILAE commissioned a Task Force to establish a practical clinical definition of epilepsy [27]. The operational definition was formulated to bring clearness and clinical relevance to the diagnostic process. Epilepsy was then considered a disease of the brain characterized by any of the following conditions:

1. *"At least two unprovoked (or reflex) seizures occurring > 24 h apart"*;

2. *"One unprovoked (or reflex) seizure and a probability of further seizures similar to the general recurrence risk (at least 60%) after two unprovoked seizures, occurring over the next 10 years"*;
3. *"Diagnosis of an epilepsy syndrome"*;

"Epilepsy is considered to be resolved for individuals who had an age-dependent epilepsy syndrome but are now past the applicable age or those who have remained seizure-free for the last 10 years, with no seizure medicines for the last 5 years."

Conceptually, epilepsy was considered a disorder rather than a disease. However, in this practical view, epilepsy was referred to as a brain disease since the term "disorder" suggests a functional disarrangement, not necessarily permanent, which minimizes the serious nature of epilepsy and is not well understood by the public [27].

Reflex epilepsies were also included in this revised practical definition. Reflex epilepsy is a disorder in which seizures are triggered by specific factors such as photic stimuli as opposed to unprovoked seizures that occur in the absence of any precipitating factors [27].

2.1.2 Classification

In 2017, ILAE presented a revised framework for Classification of the Epilepsies [1] which the principal intent was to implement a communication framework for clinical use, understandable by patients and families and largely applicable to all ages.

The classification framework comprises three distinct diagnosis levels (see Figure 2.1): seizure type, epilepsy type and epilepsy syndrome. At each stage of classification, it is also essential to identify the etiology of the patient's epilepsy as well as the presence of comorbidities, enabling early diagnosis and appropriate treatment strategies [1].

Regarding the implications for epilepsy treatment, six etiologic groups have been recognized: structural, genetic, infectious, metabolic, immune, and unknown group. Besides, some comorbidities associated with epilepsy involve learning difficulties, psychological and behavioral problems, and psychiatric features, which range in type and severity [1].

A significant clinical heterogeneity characterizes epilepsy regarding types of seizures, types of epilepsy and epilepsy syndromes. Appropriate classification is critical to adjust treatment strategies and recognize associated comorbidities.

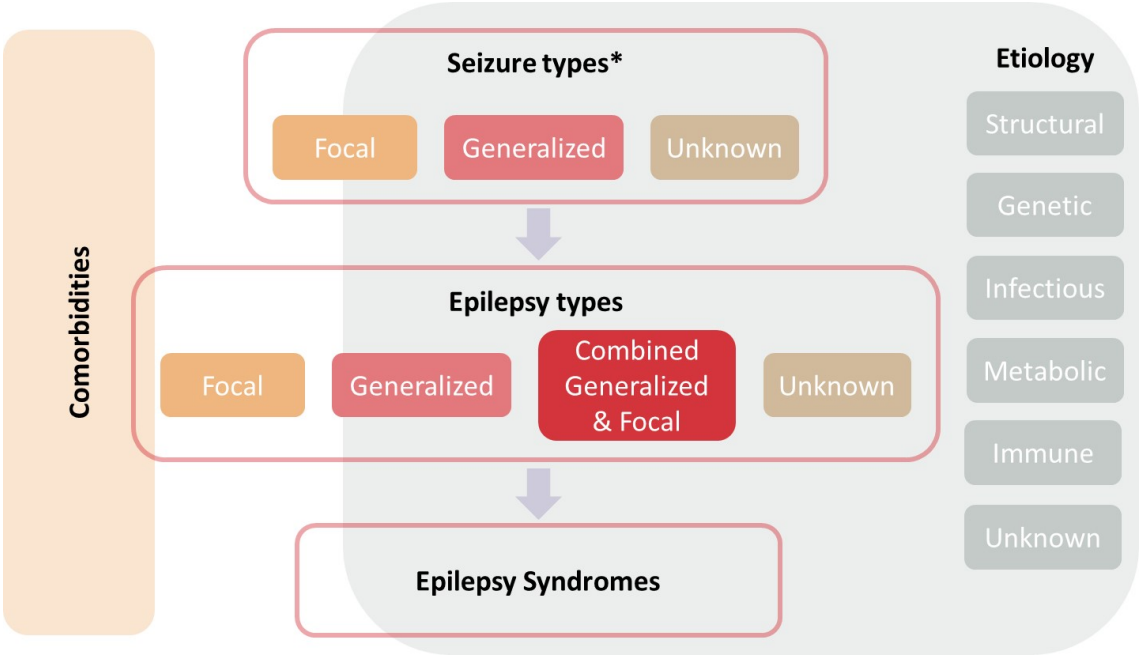

Figure 2.1: ILAE 2017 framework for classification of epilepsies. *Denotes onset of seizure. Adapted from: Sheffer et al. 2017 [1].

2.1.2.1 Seizure Type

The seizure type is the first level of the epilepsy classification framework. It requires that a previous diagnosis of an epileptic seizure has already been made.

The ILAE classification of seizure types [2] is an operational classification, not based on fundamental mechanisms. The classification chart, illustrated in Figure 2.2, is not hierarchically dependent, meaning that levels of classification can be skipped without implicating the mention of the others.

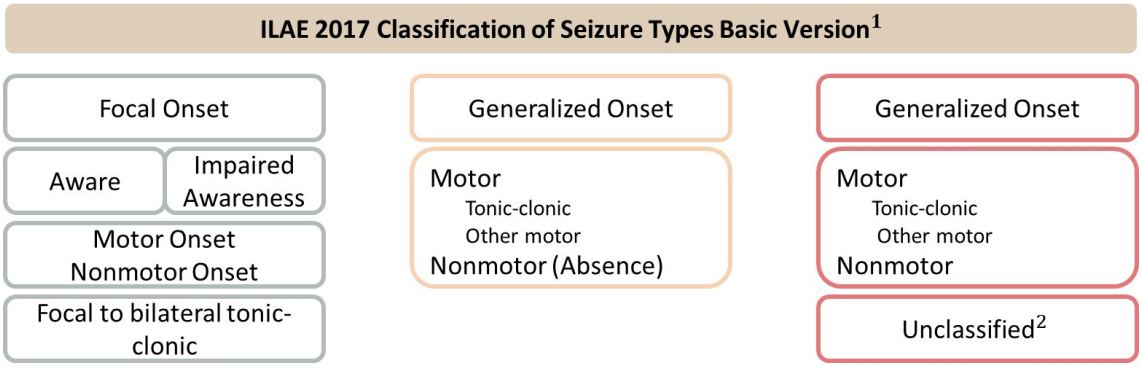

Figure 2.2: The basic ILAE 2017 operational classification of seizure types. [1]Definitions, other seizure types and descriptors are listed in the accompanying paper and glossary of terms. [2] Due to inadequate information or inability to place in other categories. Adapted from: Sheffer et al. 2017 [2].

Seizure classification starts with determining its onset: focal, generalized or unknown. Focal seizures involve networks restricted to just one hemisphere of the brain, while generalized seizures arise in both hemispheres of the brain involving bilaterally distributed networks. When it is hard to determine the onset with a certain confidence level, the seizure type is described as an unknown onset [2, 28]. Nevertheless, unknown onset seizures may later become classified as either focal or generalized onset when more information is available.

Concerning the person's state of awareness, focal seizures can be subdivided into two distinct categories: Aware and Impaired Aware. A seizure is classified as focal onset aware seizure (FOA) when the patient is aware of themselves and the surrounding environment during the seizure, even if immobile and unresponsive. Whenever the person's level of awareness alters during the seizure, it is promptly classified as Focal Onset Impaired Awareness (FOIA) [2].

The type of prominent symptom the person experiences would help to further characterize a seizure as motor or non-motor. Motor seizures lead to a change in muscle activity, while non-motor seizures can induce modifications in one of the senses [2, 28].

A particular seizure type that should be noticed is focal to bilateral tonic-clonic that begins in a particular area on one side of the brain and spreads to engage both sides. These seizures present muscles stiffening (tonic) and jerking movements (clonic) [2, 28].

Often it is impossible to classify a seizure at all, either because of insufficient information or the uncommon nature of the seizure. In these cases, the seizure is called an unclassified seizure [2].

Furthermore, it is frequent and pertinent to analyze and specify other factors in order to provide more context information. Therefore, common procedures include inspecting the seizure onset localization in brain lobes and hemispheres or registering the patient's vigilance state during the seizure.

2.1.2.2 Epilepsy Type

The second level of the classification framework is the epilepsy type. It presumes the patient has an epilepsy diagnosis as stated in the 2014 practical definition [27]. The determination of the epilepsy type is based on clinical grounds and supported by EEG findings. As shown in Figure 2.1 the epilepsy types are organized into four distinct classes [1]:

- Focal Epilepsy: includes unifocal and multifocal disorders and seizures involving one hemisphere. The interictal EEG generally displays focal epileptiform

discharges. A variety of seizure types can be presented, including focal aware seizures, focal impaired awareness seizures, focal motor seizures, focal non-motor seizures, and focal to bilateral tonic-clonic seizures.

- Generalized Epilepsy: includes a range of seizure types such as absence, myoclonic, atonic, tonic, and tonic-clonic seizures. The EEG is typically characterized by generalized spike-wave activity.
- Combined Focal and Generalized Epilepsy: includes both focal and generalized seizures and is characterized by generalized spike-wave and focal epileptiform discharges.
- Unknown: represents the cases when it is difficult to determine whether the epilepsy type is focal or generalized due to the insufficient available information.

It is relevant to emphasize the complexity of an epilepsy type since each category comprises multiple seizure types [1].

2.1.2.3 Epilepsy Syndrome

The third level of the classification corresponds to an Epilepsy Syndrome diagnosis. An epilepsy syndrome is described by a cluster of features, including seizure types, EEG findings, and imaging features which tend to occur together. Likewise, features such as age at onset and remission, seizure triggers and diurnal variation are regularly considered. Despite the presence of well-recognized syndromes, the ILAE hasn't established a formal classification of syndromes [1].

Determining epilepsy syndromes often provides information on what medications or other treatments will be most appropriate. It also may help predict if the seizures will go into remission [29].

Temporal Lobe Epilepsy (TLE) is the most common epilepsy syndrome as well as the most usual form of focal epilepsy. There are two types of TLE:

- Mesial temporal lobe epilepsy: involves the medial or internal structures of the temporal lobe. Seizures often begin in the hippocampus or surrounding area.
- Neocortical or lateral temporal lobe epilepsy: involves the outer part of the temporal lobe.

This epilepsy syndrome often appears around age 10 or 20, and it is difficult for TLE patients to become completely seizure-free with medicines, thus making surgery the best option. Therefore, usually seizure prediction research focus on these patients [30, 31].

2.1.3 Seizure clusters

Seizure clusters are epileptic episodes characterized by an increased seizure frequency. However, there is no consensus regarding the clinical definition. While some studies describe seizure clusters as more than 2-3 seizures in 24 hours, others consider it in 6-8 hours.

Seizure clusters affect people with epilepsy regardless of their sex or age and are more frequent in patients with focal epilepsy but can also occur in generalized epilepsy. It is unclear why this phenomenon occurs. However, it could be caused by a failure to inhibit the epileptic discharge in the brain, which can occur for many reasons [32].

There is evidence that seizures that occur within minutes or a few hours are more likely to come from a concordant focus. Therefore, they may not be independent of each other.

2.1.4 Treatment

The purpose of treatment for epilepsy is to achieve epilepsy freedom without side effects as soon as possible. The first-line treatment for epilepsy is Anti-Epileptic Drugs (AEDs). Nevertheless, almost one-third of patients fail to achieve seizure control with medication alone, being considered Drug-Resistant Epilepsy (DRE) patients [13, 14].

Epilepsy surgery is a well-established treatment for DRE patients. The objective of epilepsy surgery is to remove part of the epileptogenic cortex in order to make a patient seizure-free [15].

However, only a small amount of patients are eligible for this therapy. Therefore, an appropriate pre-surgical evaluation is required to select suitable patients likely to become seizure-free after the surgical treatment. The main goal of the pre-surgical evaluation is to identify the epileptogenic zone accurately and to determine whether it can be removed without significant side effects. The referred process requires a multimodality approach wherein each modality provides unique and complementary information. The basic and additional modalities include clinical history, long-term video-EEG recording, high-resolution MRI, neuropsychological evaluation, and intracranial monitoring [13, 15].

During the pre-surgical monitoring, the patient must stay in the hospital for several days. Therefore, some strategies are used to accelerate the process to reduce patient discomfort and hospital costs. Some techniques, such as medication tapering and sleep deprivation, are employed to increase the seizure frequency.

For those patients that are not eligible for surgery, neurostimulation devices, dietary therapies or clinical trials of new AEDs are alternative options [13].

In the scarcity of completely controlling a patient's epilepsy, seizure prediction plays a significant role in clinical management and treatment. This approach improves the quality of life of patients who are susceptible to the sudden occurrence of seizures.

2.1.4.1 Drug-Resistant Epilepsy (DRE)

In 2009, ILAE proposed a consensus definition for Drug-Resistant Epilepsy (DRE) [33]. According to this, DRE is defined as: *"failure of adequate trials of two tolerated and appropriately chosen and used AED schedules (whether as monotherapies or in combination) to achieve sustained seizure freedom"*. Seizure freedom was also characterized as: *"freedom from all types of seizures for 12 months or three times the preintervention interseizure interval, whichever is longer"*

Patients with drug-resistant epilepsy are at a higher risk of developing various psychological problems, such as depression, anxiety, and psychosis. Furthermore, the mortality rate is 5-10 times higher than that of the general population, including accidental injury, cognitive decline and sudden unexpected death in epilepsy (SUDEP) [14, 15].

2.2 Electroencephalogram (EEG)

2.2.1 Overview

The human brain is a complex system composed of millions of interconnected neurons. They work together in a network to process and transmit information through small electrical impulses. Therefore, it is possible to understand the brain system from the measurements or signals obtained from it [3].

The EEG measures and records the brain's electrical activity. It is a representation of voltage variations in space and time. The electrical potentials are derived from the summation of excitatory and inhibitory postsynaptic potentials generated by pyramidal cells in the cerebral cortex. To generate a large enough signal to register on an EEG, thousands of neurons will have to be in synchrony [3, 34, 35].

This medical test reflects the global dynamics of the brain's electrical activity over time. Therefore, the EEG is a valuable tool in the study and diagnosis of several abnormalities related to the improper function of the brain, like epilepsy. Furthermore, this signal is capable of capturing fast changes in the brain due to its

high temporal resolution making it a relevant tool for the research fields of predicting and detecting seizures [3, 36].

Commonly, the potentials recorded with EEG can be classified into two distinct categories (see Figure 2.3): oscillations and transient events. Oscillatory activity is characterized by sustained rhythmic fluctuations or repetitive patterns in electrical brain activity. In contrast, transient phenomena are expressed by sharp brain waves lasting only one or two cycles. In turn, each referred type can be divided into normal and abnormal activity [34]. An EEG waveform is considered abnormal when it exhibits unusual characteristics which do not correspond to the person's state of awareness, age, and other factors [37].

The waveform frequency defines normal oscillations, so they are categorized into different bands: delta (2-4Hz), theta (4-8Hz), alpha (8-13Hz), beta (13-30Hz), and gamma (>30Hz).

Abnormal oscillations are divided into seizures, and burst-suppression [34]. The EEG pattern described by the term burst-suppression consists of a continuous interchange between high-voltage slow waves and suppressed electrographic activity. This motif is noticed in several conditions such as deep coma, drug intoxication, and encephalopathies [38]. Regarding seizures, few or many EEG channels can present synchronization, hyperexcitability and oscillations. However, these trends are merely representative, often identified as a result of a pronounced change from background activity. Indeed, there is no single 'epileptic' EEG: fluctuations between patients, between seizures in the same patient and within a single seizure are observed [3].

Normal transients cover a range of sleep potentials, as well as a variety of artifacts, [34]. Sleep potentials may occur before and after the switch of the subject state of alertness, essentially in stage 2 of NREM sleep [37]. Artifacts are noncerebral electrical potentials that are detected with EEG. They are defined as noise and can be physiological, including eye blinks, cardiac impulses, breathing, chewing, muscle activity, or can be external as electromagnetic interference from the surrounding environment and incorrect electrode placement [3, 34].

Abnormal EEG transients can be classified as epileptiform or non-epileptiform potentials. The first transients are essential for the diagnosis of epilepsy, while the second are indicators of several other encephalopathies [34].

As depicted, EEG is a medical tool widely used to detect and analyse epileptic seizures. However, the non-linear and non-stationary nature of the EEG and the presence of artifacts that are difficult to remove without affecting the measurement of true brain activity make it a complex signal. Consequently, understanding the morphology of the signal and its manual inspection are challenging [3, 39].

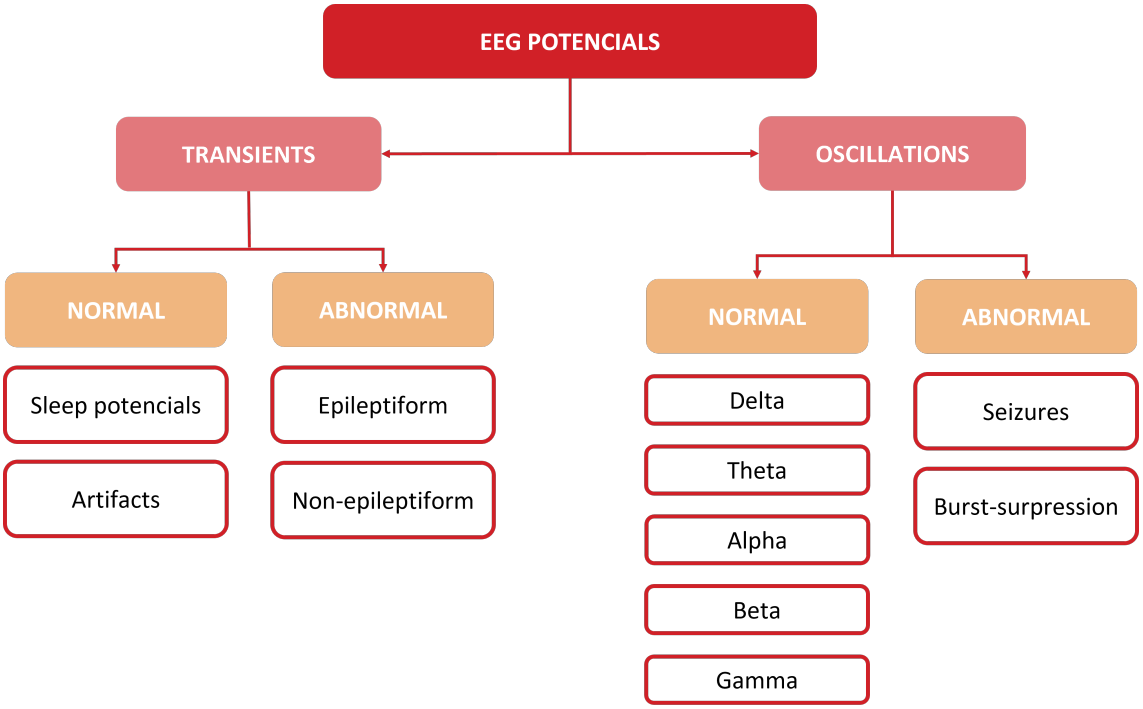

Figure 2.3: Categorization of EEG activity.

2.2.2 Signal acquisition

Besides its high temporal resolution, the EEG has poor spatial resolution because records are restricted by the number and the placement of the electrodes and the properties of the head.

The acquisition of the EEG signal can be made at different spatial scales. It is possible to collect macroscopic measurements either non-invasively from the scalp (Scalp EEG) or over surgical procedures that allow intracranial recording (intracranial EEG (iEEG)). Both present advantages and disadvantages [3].

Scalp EEG

As referred above, scalp EEG signals are noninvasively acquired through electrodes placed on the scalp. In order to reduce the impedance, an electroconductive gel is used [36]. The signal must spread over a range of nonneuronal layers like cerebrospinal fluid, skull, and scalp. As a result, scalp measurements are attenuated, requiring much larger regions to be actively synchronized for an EEG signal to record [3].

Standard systems for electrode placement are frequently used to make records easily comparable between and within patients. This system is denominated as the International 10-20 System. A ground and system reference electrode and 19

recording electrodes are placed on the scalp area according to a computed percentage of standard distances. In particular, they are situated at 10%, 20%, 20%, 20%, 20%, and 10% from nasion to inion.

As depicted in Figure 2.4, electrodes are denoted with the prefix "F", "Fp", "C", "T", "P" and "O" accordingly to the electrode placement over frontal, frontalpolar, central, temporal, parietal, or occipital regions, respectively. The assigned prefix is followed by an odd number for electrodes placed on the left, an even number for right electrodes or the letter "z" for midline electrodes [3, 34, 36].

All EEG recorded signals are organized into channels that measure the voltage difference between two electrodes. Electrodes can be arranged into either bipolar or referential montages, depending on how a particular EEG electrode is referenced. EEG signals may present different morphologies relying on the chosen montage [34].

Bipolar montages are frequently organized in straight-line chains of electrodes, where the second input to each channel is the next electrode in the line. Therefore, each channel measures the voltage difference between two adjacent electrodes. Alternatively, in referential montages, the second input to each channel is a reference electrode placed either on the scalp or in other body parts like ear lobes, mastoids or nose. Clinically, one of the most used montages for scalp EEG recordings is the anterior-posterior longitudinal montage, so-called "double banana", represented in Figure 2.4 b) [34].

iEEG

Regarding iEEG, electrodes are implanted directly on the brain through a surgical procedure. Electrodes can be placed on the exposed brain surface to record the electrical activity of the cerebral cortex or into subcortical systems to record more profound brain activity [3, 36].

There are two main types of invasive electrodes: strips or grids and depth electrodes (see Figure 2.5). Subdural strips or grids are placed on the brain surface and record electrical activity from many points, with the grids covering larger areas. On the other hand, depth electrodes are thin wires that look like a needle and contain several recording sites along each electrode. They are used if seizures may arise in deeper brain areas instead of on the surface [40].

Intracranial records are often acquired for pre-surgical evaluation to determine regions of the brain to be resected when non-invasive procedures cannot localize the seizure onset zone [3]. The majority of time series databases used in research of seizure prediction arise from this context [34].

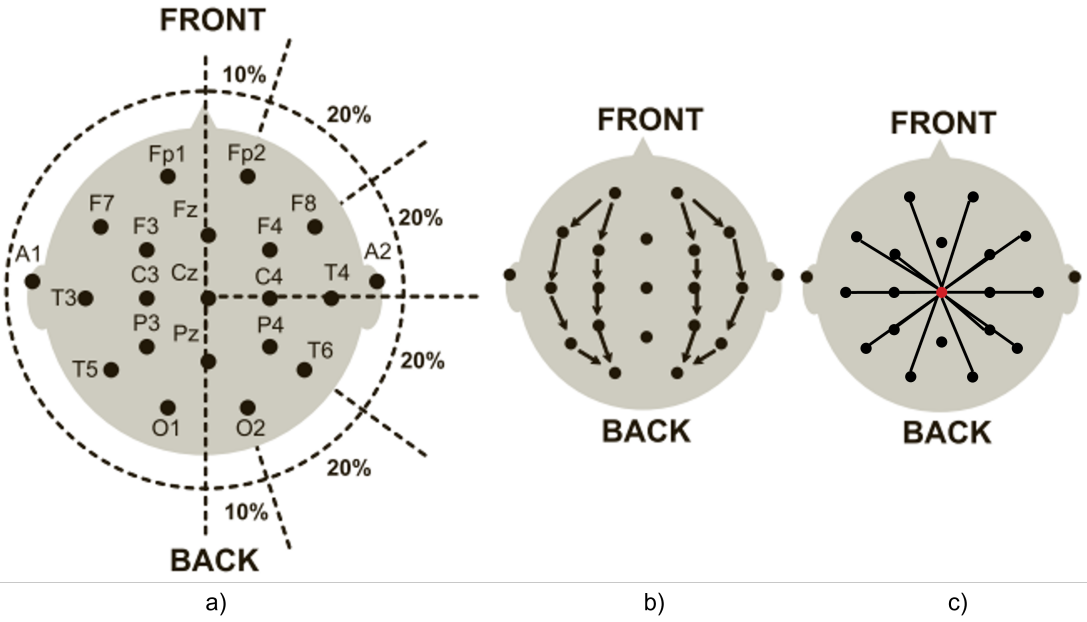

Figure 2.4: International 10-20 system for placement of scalp EEG electrodes. In (a) is shown the standard positions and names of the electrodes. In (b) is represented a bipolar montage and in (c) a referential montage. Adapted from: Varsavsky et al. 2011 [3].

Scalp EEG *vs.* iEEG

Noninvasive EEG recording is an easy and inexpensive diagnostic tool that is frequently capable of supplying relevant information to physicians.

The scalp EEG signal also bears the advantage of covering a larger area than the invasive one. However, this procedure has a range of limitations compared with intracranial records, including extracranial artifacts, the inefficiency to accurately register part of the activity in beta and gamma bands, and the inability to record activity from the deep brain [3, 34]. Despite being less invasive, surface recordings would arise constraints on the patient due to the inconvenience of wearing an EEG cap if long acquisition periods are necessary [18]. Scalp-recorded EEG often constitutes a preliminary step to more detailed intracranial EEG due to its inconclusive results [3].

Intracranial EEG recordings have less contamination with artifacts and more excellent proximity to onset zones, resulting in a higher signal-to-noise ratio and a better spatial resolution [18, 34]. Nevertheless, the invasive recording presents a superior risk of infection or hemorrhage [34]. Such a procedure also fails to meet a standard for electrode placement since each decision is made on a patient basis [3].

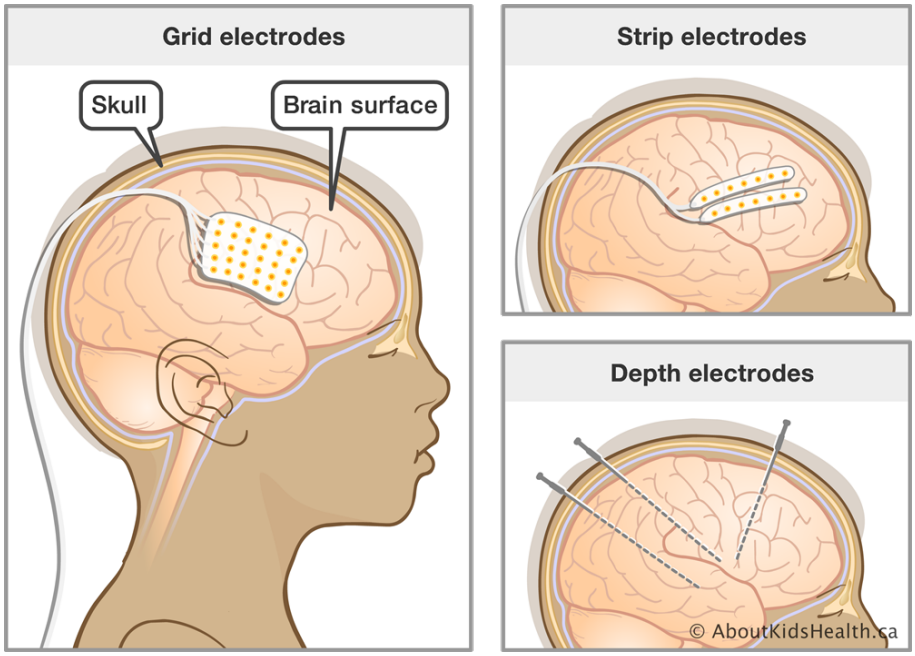

Figure 2.5: Different types of invasive electrodes. Extracted from: [4].

2.2.3 EEG seizure period division

In addition to differentiating epileptic patients from non-epileptic ones, EEG signals also help distinguish the different seizure stages due to their patterns (see Figure 2.6). Hence, a typical EEG record from an epileptic patient can be divided into different periods: preictal (the period preceding the seizure), ictal (the period corresponding to the seizure), postictal (the period after the seizure), and interictal (the period between the postictal and preictal stage of two consecutive seizures) [39].

It is worth noting that the preictal state is the most difficult one to determine and manually annotate by experts since it is associated with substantial heterogeneity. It can be justified by the diversity of seizure types and the presence of discrepant onset mechanisms, resulting in distinct and complex preictal dynamics. This diversity is present between patients as well as within seizures from the same patient. Since the preictal state plays a significant role in seizure prediction, the complex nature associated with this state represents a major challenge for ongoing research [19, 39].

2.3 Seizure Prediction

The unpredictability of seizures is the main problem for patients with uncontrollable epilepsy and their families. Predicting epileptic seizures would effectively

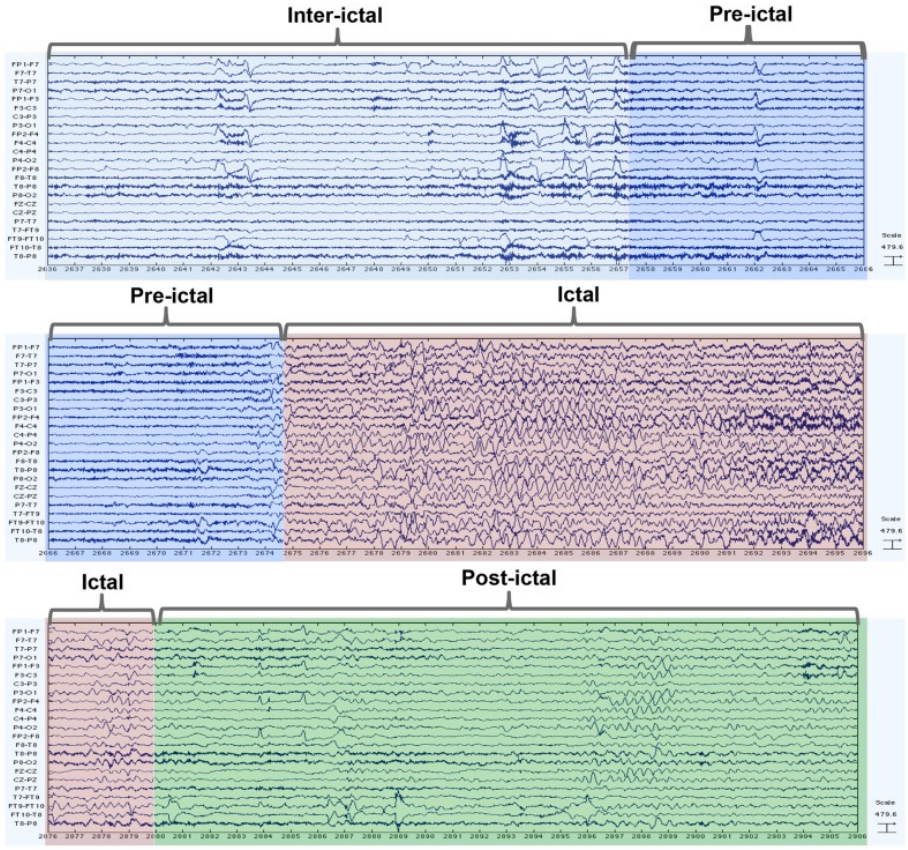

Figure 2.6: Different periods of an epileptic seizure annotated on the EEG signal. All four states of ictal, preictal, ictal, postictal and interictal are colour coded. Source: Moghim et al. [5].

improve their quality of life and safety. An accurate prediction at an early enough stage before seizure onset would provide new therapeutic options such as warning devices that enable the patient to avoid dangerous situations or even intervention devices capable of controlling the seizure by delivering anticonvulsive drugs or triggering electric stimuli [6].

A seizure prediction algorithm should be capable of anticipating an epileptic seizure by raising an alarm before the seizure onset. The goal is to construct a system able to read online data and adequately notify the patient regarding a seizure that will arise on a well-defined occurrence period with a predefined horizon, which must allow enough time to take action. Ideally, this should minimize the unpredicted seizures and the false alarms in order to reduce interruption to an individual's life [6, 16].

Seizure onset

A critical point in seizure prediction is determining the beginning of a seizure. The onset time can be electrographic or clinical. The first type is detected from the first clinical symptoms, whereas the second is determined from the first visible variations in the EEG. Considering clinical signs are often imperceptible and difficult to recognize, especially in FOIA and non-motor seizures, and the EEG onset anticipates the clinical onset, it is practical to estimate the beginning time of a seizure from the electrographic onset [18, 41].

Lead seizure

In prediction studies, the seizures used must be independent events. Therefore, it is necessary to manage seizure cluster episodes. When this phenomenon is present in the data used in the study, the authors only use the first seizure of the cluster, usually known as the lead seizure.

Since there is no agreement concerning the definition of a seizure cluster, a range of values has been used in the literature as the minimum seizure-free interval to consider a continuous set of seizures independent of each other.

Seizure detection

Seizure detection is a parallel research field whose algorithms aim to detect the electrographic seizure onset, which may arise a few seconds before the first clinical signs. In opposition to seizure prediction, which intends to identify the preictal state sufficiently early before the EEG onset, seizure detection does not provide enough time to take action. Aside from warning the patient that a seizure is up-coming, this approach may supply clinicians with detailed seizure information valuable for epilepsy management. Furthermore, when implemented within closed-loop intervention systems it could be effective [18, 41].

Albeit seizure detection algorithms are superior in terms of performance, seizure prediction algorithms are preferred for more promptly responses [17].

Seizure forecasting

Seizure forecasting is also an emerging parallel research field which aims to determine periods of a high probability of seizure occurrence. As an alternative to seizure prediction, which aims to distinguish between interictal and preictal periods, seizure forecasting identifies the brain state in which a seizure is most likely to occur,

the proictal period. The proictal state relates closely to cycles connected to EEG activity, such as the circadian and multi-day rhythms [42].

2.3.1 Seizure prediction characterization

The preliminary work on the seizure prediction field dates back to the 1970s. Since then, several algorithms to predict seizures using EEG data have been proposed, showing promising results. Nevertheless, in the early 2000s, some publications found a less promising performance of the developed algorithms than those previously reported [18].

Subsequently, in 2003, Winterhalder et al. [6] proposed the "seizure prediction characteristic" to evaluate and compare the different seizure prediction methodologies founded on clinical, behavioral, and statistical considerations. Therefore, two crucial concepts were introduced: Seizure Prediction Horizon (SPH) and Seizure Occurrence Period (SOP).

Ideally, a seizure prediction method would announce the exact onset time of a seizure. However, some uncertainty is expected with EEG signal-based prediction methods. Thus, it was suggested the use of SOP, defined as the time interval during which the seizure is presumed to arise. Furthermore, to make any intervention feasible, the existence of a minimum time interval between the raising of the alarm and the beginning of the SOP is crucial. This time window is designated as SPH, also known as Intervention Time (IT).

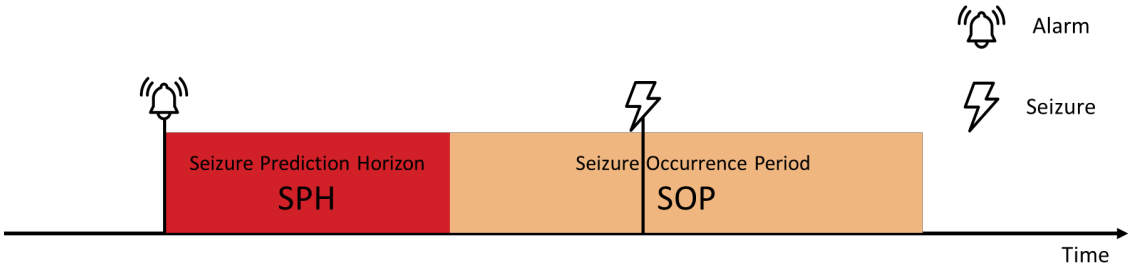

Figure 2.7: Visual representation of SPH and SOP. Adapted from: Winterhalder et al. [6].

No optimal values have yet been found for SOP and SPH. Although, their choice should be reasonable regarding patient and clinical considerations.

On the one hand, the SPH values should provide enough time to take preventive actions. Whereas intervention devices may need only a few seconds to control the imminent seizure, warning devices have to raise the alarm minutes before the seizure onset, providing more time to avoid dangerous situations.

On the other hand, the SOP values can range from minutes to hours. Interventions like electrical stimulation and anticonvulsive drug-delivering should last the whole seizure SOP. Therefore, longer window times may bring undesirable consequences for such long interventions. Furthermore, in the case of warning systems, longer SOPs increases the patient's anxiety.

The values adopted for both parameters play an essential role in the performance of the prediction algorithms.

2.3.2 Performance evaluation

The performance of a given algorithm is commonly evaluated by some metrics such as Sensitivity (SS) and Specificity (SP). The confusion matrix is typically used to define such measures in standard machine learning problems [43]. Considering seizure prediction as a binary classification problem, where the positive class (class 1) is the preictal period and the negative class (class 0) is the interictal period, it is possible to define the following confusion matrix (Table 2.1) regarding clinical and predicted labels.

Table 2.1: Confusion matrix for evaluation of sample performance in Machine Learning problems.

		Clinical Label	
		Preictal	Interictal
Predicted Label	Preictal	TP	FP
	Interictal	FN	TN

While sensitivity expresses a classifier's effectiveness in identifying the positive labels, specificity characterizes how effectively a classifier identifies the negative ones [43]. These metrics are calculated by Equations 2.1 and 2.2, respectively.

$$SS_{sample} = \frac{TP}{TP + FN} \tag{2.1}$$

$$SP_{sample} = \frac{TN}{TN + FP} \tag{2.2}$$

The metrics are computed based on sample classification, not offering information regarding the number of correctly predicted seizures or false alarms. Thus, they were adapted to be more informative in the seizure prediction field.

Therefore, the performance of a seizure prediction algorithm is frequently assessed by Sensitivity and False Positive Rate per Hour (FPR/h) [6, 17, 34]. These

measurements require the correct distinction between true and false alarms.

As illustrated in Figure 2.8, an alarm is expressed as a true alarm when the seizure onset occurs during the established SOP. In opposition, if the seizure arises outside the SOP time, the raised alarm is considered a false alarm [6].

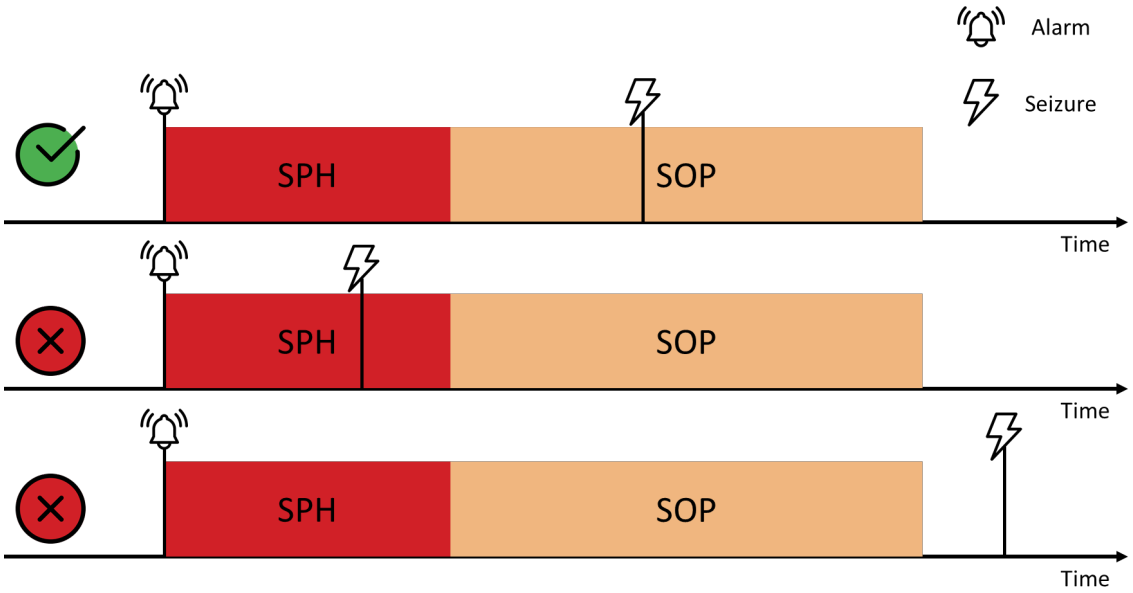

Figure 2.8: Visual representation of true and false alarms in seizure prediction, considering SPH and SOP.

Concerning the performance indices, sensitivity is the most adopted metric in seizure prediction studies and measures the fraction of the correct predictions for all seizures, as described by Equation 2.3 [6, 34].

$$SS = \frac{Predicted\ seizures}{All\ seizures} \tag{2.3}$$

In turn, FPR/h represents an appropriate metric for specificity in the field of seizure prediction, measuring the occurrence of false alarms during an hour [6, 17]. Furthermore, this metric gives more information to clinicians than specificity. Once FPR/h translates the number of false alarms per hour, the clinician can adapt the system to the most suitable application regarding the psychological and mental resilience of the patient. For instance, a system with a high value of FPR/h would not be appropriate for interventions like anticonvulsive drug-delivering or electrical stimulation since it must not be supplied several times in an hour. Thus the applicability of the prediction systems would not be interpreted by the clinicians if the specificity metric was used.

However, there is no agreement regarding its definition. In some studies, FPR/h is determined by dividing the number of false alarms by the total duration of the

analyzed recording, as described by Equation 2.4. Although, it must be taken into account that there is a period during which the raised alarms are counted as true, where false predictions cannot arise by definition. Furthermore, when an alarm is fired, there may be a period during which it is impossible to raise a new alarm. This time interval is called a refractory period and is equal to the sum of SOP and SPH duration. Consequently, in other studies, the concept of corrected FPR/h is used, taking into account the considerations mentioned above. Thus, it is defined as the proportion between the number of false alarms and the period during which alarms can actually be raised, as expressed by Equation 2.5 [18].

$$FPR/h = \frac{False\ alarms}{Total\ time\ analyzed} \tag{2.4}$$

$$FPR/h = \frac{False\ alarms}{Interictal\ duration - False\ alarms \times (SOP + SPH)} \tag{2.5}$$

As expected, the ideal scenario would be where all seizures were correctly predicted and no false alarms were raised, expressed by a performance value of 100% sensitivity and 0 FPR/h.

In order to approximate this utopian goal, a trade-off between both metrics should be taken into account since they present a close interdependency relation: an increase in sensitivity leads to an increase in FPR/h [6].

Excessive false alarms may lead to undesirable consequences of unnecessary interventions and patients' distrust of the warning systems. As a result, clinical applications performing a high sensitivity at the cost of a high FPR/h are doubtful regarding the patient's quality of life. For this reason, a maximum value for false positive rate must be defined, respecting the patient, chosen intervention system and clinical considerations [6, 34].

A reasonable value for FPR/h_{max} may be based on the average seizure incidence. On the one hand, in pre-surgical monitoring, the seizure frequency is uncommonly elevated in response to the reduction of AEDs administration. The maximum average frequency reported is 3.6 seizures per day (0.15 seizures per hour). On the other hand, DRE patients, under normal circumstances, have an average frequency of around three seizures per month (0.0042 seizures per hour) [6].

Considering the majority of available databases of epilepsy recordings are from pre-surgical monitoring data, an expected reasonable value for FPR/h_{max} would be 0.15. It indicates that, even if all seizures can be correctly predicted, 50% of the fired

alarms would be deemed false alarms for patients during monitoring. Although, this value increases to 97% for epileptic patients under normal conditions, which would be improper [6].

To sum up, the performance of a seizure prediction algorithm is characterized by the dependence of the Sensitivity (SS) on the False Prediction Rate (FPR/h), Seizure Occurrence Period (SOP), and Seizure Prediction Horizon (SPH) [6].

2.3.3 Statistical Validation

Along with the characterization and performance assessment, statistical validation plays a fundamental role in the evaluation of seizure prediction algorithms as well as in the comparison of distinct prediction models. It should be implemented to confirm if the developed algorithms perform above chance level and identify the preictal stage [6, 17, 18].

For the purpose mentioned above, several techniques have been suggested: comparisons with random and periodic predictors, analytical random and baseline predictors, comparisons with chance level by analyzing the areas under the receiver operating curve (AUC), surrogate methods, and non-parametric methods [17]. The most used methods are explained in the following subsections.

2.3.3.1 Analytical random predictor

Random prediction is one unspecific method in which alarms are raised completely randomly without any information included in the EEG data [6, 7, 44].

Schelter et al. [7, 44] presented an analytical random predictor founded on a homogeneous Poisson process for false predictions. Hence, the probability of raising an alarm at each sampling point extracted from a times series is defined by Equation 2.6.

$$P_{\text{Poisson}} = \frac{N_{\text{false alarms}}}{N_{\text{samples}}} \tag{2.6}$$

Considering now a period of duration equal to SOP, the probability of an alarm being triggered within this interval for a given FPR/h can be approximated by Equation 2.7. Although, this approximation is only valid if the product of the maximum false prediction rate by the occurrence period ($FPR/h \times SOP$) is significantly smaller than one, guaranteeing that the patient is not under constant warning.

$$P \approx 1 \; - \; e^{-FPR/h \times SOP} \approx FPR/h \times SOP \tag{2.7}$$

The probability P (Equation 2.7) constitutes the sensitivity of a random prediction method, since it is defined as the probability of raising at least an alarm during the SOP [6]. Furthermore, it comprises the basis for a significance level to test whether the sensitivity of a prediction method is better than a random one.

Nevertheless, other factors must be taken into account in the significance level: several seizures are usually investigated, and the dimension d of the extracted features vector has to be considered since increasing the number of predictors increases the probability of predicting seizures by chance. However, in ML models, a single-dimension output (d=1) can be obtained even if multidimensional inputs are used.

Hence, the probability of randomly predicting at least k of K seizures can be determined by the binomial distribution defined in Equation 2.8.

$$P_{\text{binom}}(k,K,P) = 1 - \left[\sum_{0}^{j \leq k} \binom{K}{j} P^j (1-P)^{K-j} \right]^d \tag{2.8}$$

Lastly, the critical value of sensitivity for a random predictor, considering a specific significance level α, is given by Equation 2.9.

$$\sigma_{\text{rand}} = \frac{argmax_{\text{k}}\{P_{\text{binom}}(k,K,P) > \alpha\}}{K} \times 100\% \tag{2.9}$$

The main advantage of this statistical validation method is based on the analytic expressions for its sensitivity which is simple and easy to apply. The presented equations would provide information regarding the minimum number of seizures that must be used to guarantee that performance above the chance level can be verified. However, the analytical random predictor is implemented following the homogeneous Poisson distribution for false predictions, which may not be appropriate. Moreover, such a method might be too conservative from the statistical point of view and slightly less powerful.

2.3.3.2 Surrogate seizure predictor

Alternatively to analytical random predictors, the surrogate seizure predictors were suggested based on Monte Carlo simulations. They are generated by constrained randomizations of the original seizure predictor. These seizure-predictor surrogates are designed to share specific aspects with the original predictor despite being random. Accordingly, although computationally more complex, this approach presents greater flexibility than analytical random predictors as it allows one to test several null hypotheses based on particular assumptions and constraints. Therefore, if the performance of the original seizure predictor outperforms the predictor sur-

rogates, the underlying null hypothesis can be rejected. In this case, it means that the developed algorithm performs above chance level [7, 17].

Andrzejak et al. [7, 45] introduced the seizure times surrogates method, which replaces the original seizure times with artificial ones. As illustrated in Figure 2.9, these artificial onset times are generated by randomly shuffling the original interictal intervals, preserving the inter-seizure-interval distribution, the total number of seizures and the original measure profiles. The prediction algorithm's performance is again calculated from the maintained measure profile regarding the seizure times of the surrogates and compared with the predictive performance obtained for the original seizure times. This approach can be applied to any type of analysis (algorithmic or statistical), providing a high confidence level [7, 18]. Although, the practical application of seizure-times surrogates can be challenging since only a few seizures are sometimes included in the EEG recordings, which can also contain gaps. These drawbacks can make it impossible to generate sufficient independent surrogates in order to obtain significance [7].

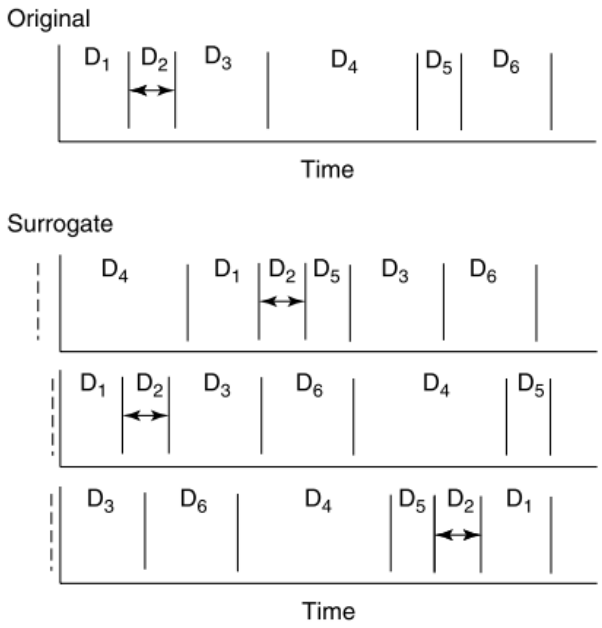

Figure 2.9: Representation of the original seizure times and the surrogate times bootstrapped from the inter-seizure intervals. The arbitrary onset times for the surrogates are originated through a uniform distribution and are represented by the dashed vertical lines. Source: Schelter et al. 2008 [7].

2.3.4 Concept drift and class imbalance

In the real world, machine learning models deal with concepts of interest dependent on some hidden context, not evidently present in predictive features. Variations in the hidden context are frequent, leading to changes in the target concept or the underlying data distribution. This problem is generally known as concept drift and makes challenging the task of learning a model from data [20].

Concerning epilepsy, the referred variations comprise alterations in the brain dynamics depending on exogenous and endogenous factors, such as changes in behavior and mood, cognitive disturbances, circadian rhythm (sleep-wake cycle, time of the day, week, month and year), medication, and others [19]. Medication is a relevant factor since the available databases are essentially constituted by pre-surgical monitoring data, in which patients are deprived of AEDs. As a result, an artificially high seizure frequency is reported, which may not represent a common event. Despite the medication tapering, its effect does not cease instantly, and the seizures become more frequent over time. Furthermore, the activities of patients under pre-surgical monitoring may be quite different from those of regular daily routine since they are mostly seated or lying down. This altered state, along with sleep deprivation, may promote notable variations in the data distribution, affecting the efficiency of the trained prediction models [6, 19, 46].

Class imbalance is another critical issue in machine learning, characterized by the underrepresentation of one class, generally the positive one, in the dataset. It may lead to bias in the learning algorithm towards the majority class and challenging interpretation [21].

In the context of seizure prediction, seizures are relatively rare events. The interictal period (the negative class) is substantially more extended than the preictal (the positive class). This issue may induce a specialization of the classifier over the interictal class [17].

It is necessary to build robust models capable of handling changes in concepts over time and data imbalance to overcome the referred problems. The implementation of ensemble techniques and the inclusion of exogenous variables have proven to help with this purpose [20, 21].

2.4 Explainability

Throughout the last decades, the appearance of more powerful computing systems, advanced learning algorithms and accessible and more extensive databases has

led to remarkable improvements in the field of Machine Learning (ML) [8]. Consequently, it has been extensively adopted in different areas such as financial systems, advertising, marketing and medicine [47].

However, along with the increased efficiency of the models, their complexity has also grown. Therefore, understanding the mechanisms and rationale behind their decisions has become challenging, and their predictions difficult to interpret [8].

Indeed, a definite trade-off between the efficiency of an ML model and its capacity to produce explainable and interpretable decisions is evident, as illustrated in Figure 2.10. On the one hand, there are the commonly described white-box models whose results are easily interpretable and include linear and decision tree-based models. However, their accuracy is outperformed by the black-box models, which are complex machine learning models such as Support Vector Machines, Random Forest, and Neural Networks [8, 47].

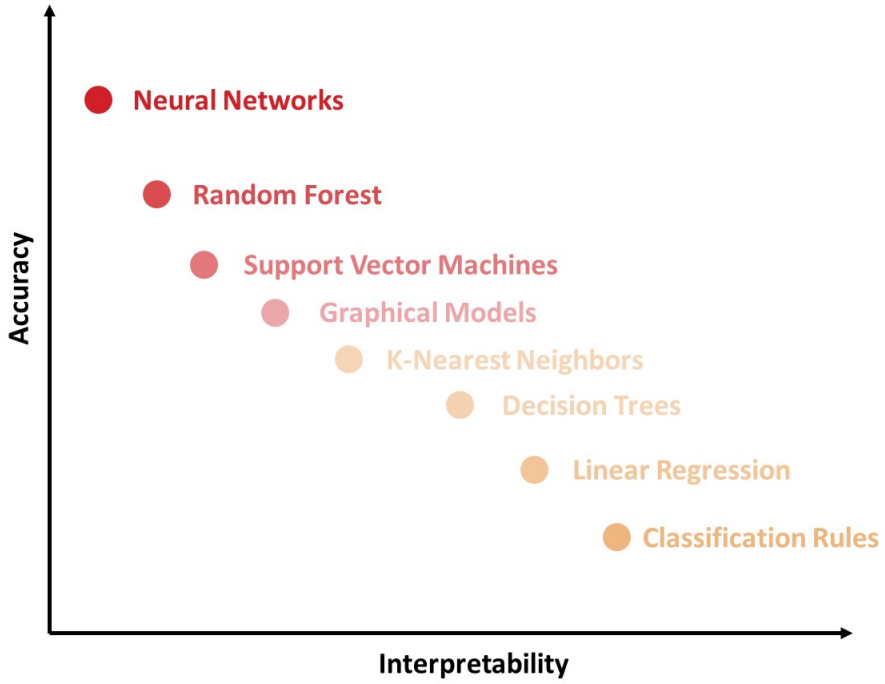

Figure 2.10: The trade-off between interpretability and accuracy of some relevant ML models.

Machine learning users have difficulty trusting complex systems whose decisions cannot be well-interpreted. It happens mainly in areas where moral and fairness issues are naturally present, such as healthcare or self-driving cars. The deployment of ML models in such sectors has led to an increased interest in optimized systems not only at the performance level but also on the level of other essential criteria, including safety, trustworthiness, fairness, robustness, and the right to explanations

[8, 9].

Subsequently, it contributes to the arrival of the field of explainability, an area responsible for understanding and interpreting the ML systems behavior [8].

Changes in policy, law and regulation are also responsible for the increased interest in this field [48]. In 2018, the European Union's General Data Protection Regulation (GDPR) forced the industries to explain any automatic decision-making process:*"a right of explanation for all individuals to obtain meaningful explanations of the logic involved"* [23].

Explaining the real world involves many layers before it reaches humans (see Figure 2.11). The world is captured by collecting data and is abstracted by learning models to predict data for a specific task. Explainability is the last layer that helps humans to understand the model and its decisions [10].

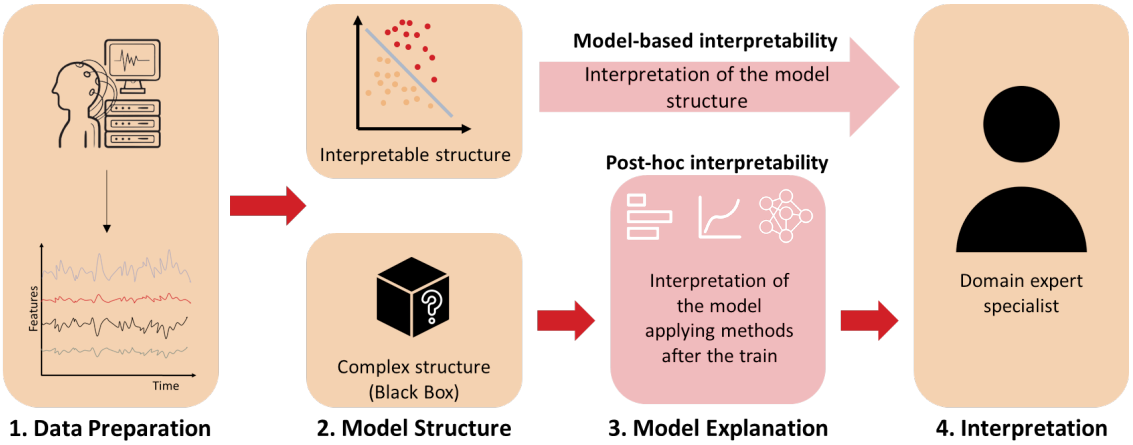

Figure 2.11: The big picture of explainable machine learning.

Explainability is still a very recent field of investigation. Therefore, there is no consensus regarding many aspects, namely, how to organize and evaluate the different explanations and define the terms of interpretability and explainability.

Interpretability and Explainability

The terms interpretability and explainability are closely related and are usually used interchangeably. They lack a formal and rigorous mathematical definition, although some non-mathematical descriptions have been proposed [8, 10]. Despite not being the definitive ones, these definitions will be used throughout this thesis.

One of the most accepted definitions of interpretability is the one presented by Doshi-Velez et al. [9], who defined it as *"the ability to explain or to present in understandable terms to a human"*. In turn, Miller et al. [49] have also proposed

a definition for interpretability: *"the degree to which a human can understand the cause of a decision"*.

Based on the definitions mentioned above, interpretability is essentially related to identifying the cause-effect relationship behind the outputs of a model. Its goal is to describe the internals of a system in a human-understandable way. The higher the interpretability of a system, the easier it is for humans to predict the model's output based on the data input [8, 10, 22].

On the other hand, explainability is associated with the model's capacity to present the reasons for its behaviour and provide insights about the causes of its decision in human terms without requiring a complete understanding of its internal mechanisms. [48].

Some authors defend that interpretability alone is insufficient and that the presence of explainability is also crucial. An interpretable model also needs to be completed with the capacity to defend their actions and provide relevant explanations. Hence, regarding machine learning systems, interpretability does not necessarily imply explainability or vice versa. [8, 22].

Not all machine learning systems need to be interpretable or explainable. In some cases, it is not critical to understanding the reason behind a decision. It is enough to ensure that the prediction is accurate. Some models may not require explanations because they are applied in a low-risk environment, meaning mistakes have no significant consequences. Additionally, methods that have already been extensively studied and evaluated do not require explanations since it is trusted even if it is not perfect [9, 10].

Doshi-Velez et al. [9] believe that the necessity for interpretability arises from incompleteness in the problem definition, raising a potential obstacle to optimization and evaluation.

Explainable and interpretable models fight the incompleteness of machine learning systems, capturing validation and trust from the scientific community and general society. They provide explanations of their predictions which contribute to the trust problem by enabling domain experts (e.g. clinicians) to ensure that the model makes the right and the wrong predictions for the right reasons. Furthermore, proper explanations are crucial to getting insights into how this model is working and help to improve its performance [47].

2.4.1 Taxonomy

The landscape of interpretability methods is viewed from various perspectives. They can be analyzed according to diverse concepts (see Figure 2.12) [8].

A significant split of interpretability methods could happen based on the way that interpretability is achieved: by restricting the complexity of the machine learning model (intrinsic) or by applying interpretation methods after model training (post-hoc). Intrinsic interpretability is associated with white-box models, which are considered interpretable due to their simple architecture. Post-hoc methods are usually applied to more complex models, although they can also be employed to intrinsically interpretable models [8, 10].

It is also possible to distinguish the methods regarding their ability to produce explanations independently of the model (model-agnostic) or restricted to a single model or group of models (model-specific) [8, 10, 47].

Additionally, the methods can be divided according to the scale of interpretation. If the technique explains only one specific instance, it is defined as local, and if the method explains the overall model, it is described as global [8, 10, 47].

Finally, one crucial factor that should be taken into account is the different result types. It is possible to examine summaries of the features' characteristics through statistics or graphic plots. It is also possible to understand the model's internals, for example, by outputting the linear weights or the structure of the trees, as it happens with intrinsic interpretability. Furthermore, the explanations can also be produced by studying the samples of the dataset and finding specific characteristics to compare new data points to their dynamics [10].

2.4.2 Evaluation

It is unclear how to measure the interpretability in machine learning models, although there is some initial research and an attempt to formulate some procedures for evaluation [10].

Doshi-Velez et al. [9] proposed three levels to evaluate the interpretability of an ML model: application, human and function level evaluation, as illustrated in Figure 2.13.

Application-grounded evaluation involves human experiments on real applications. The most reliable way to show that a new research product will work is to have it tested by the end-user. This approach measures how efficiently human-produced explanations help other humans to complete a particular task. Application-level evaluation is the most direct way to assess the system's objective. However, it is not

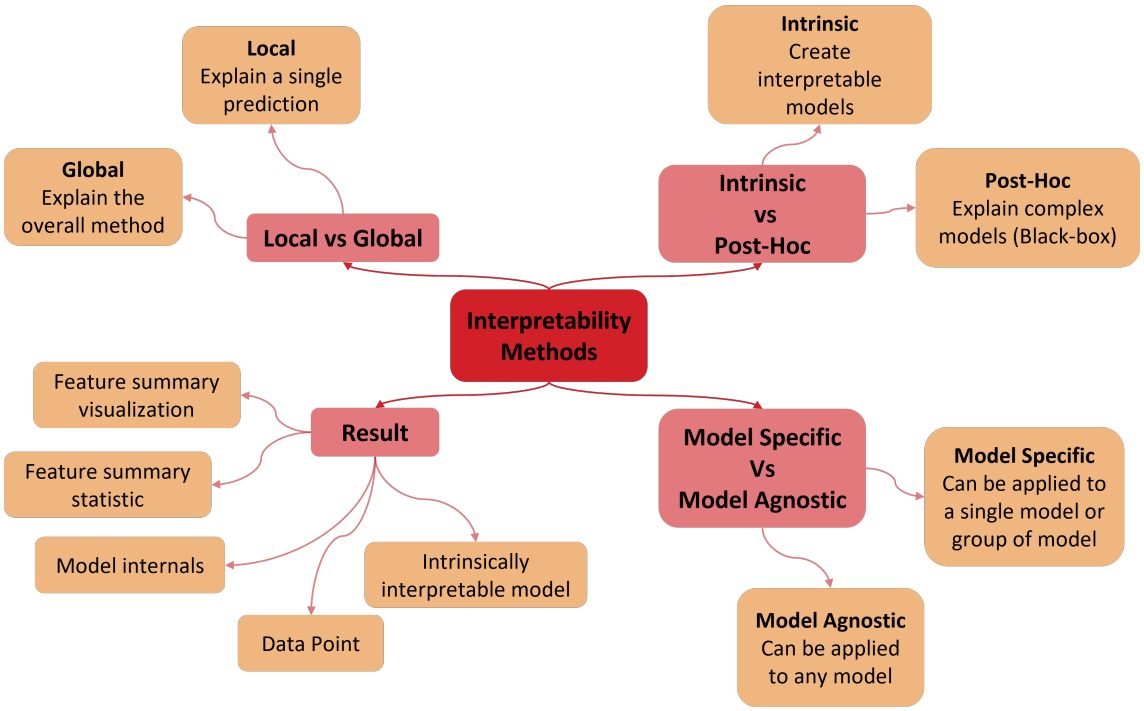

Figure 2.12: Taxonomy mind-map of Machine Learning Interpretability Techniques. Adapted from: Linardatos et al. [8].

the most accessible metric. It requires a high cost and time since these experiments are carried out with domain experts, which are challenging to get in touch with and have to be monetarily compensated for their work [9, 10, 22].

Human-grounded evaluation carries out simplified-human experiments that maintain the core of the end application. This evaluation can be accomplished by laypeople, making the experiments cheaper and easier to find testers since it does not require domain experts. The human-level assessment depends on the quality of the explanations and can be applied, for instance, by asking the testers to choose the best one between a set of explanations [9, 10, 22].

Functionally-grounded evaluation employs some formal definition of interpretability as proxies or simplified tasks to analyze the quality of the explanations. This approach is appealing because it does not involve human experiments and, consequently, does not require additional time and cost as the other evaluation levels. The function-level evaluation is more appropriate when the class of models used has already been validated by human experiments. The challenge inherent to this evaluation method is to choose the most suitable proxy [9, 10, 22].

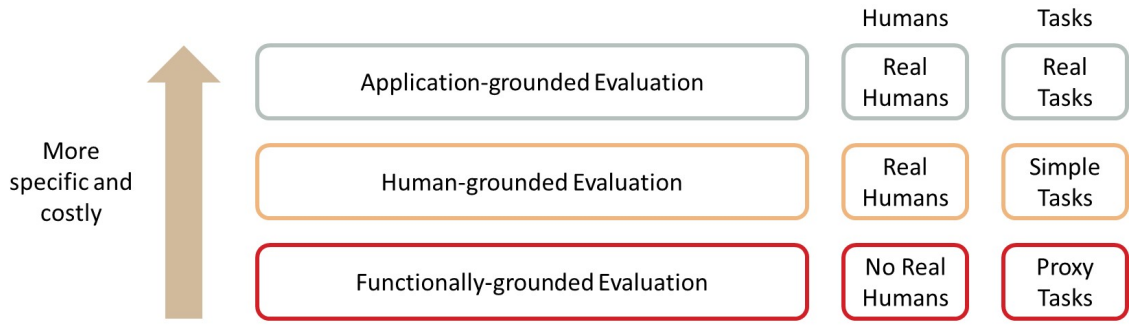

Figure 2.13: Taxonomy of evaluation approaches for interpretability. Adapted from: Doshi-Velez et al. [9].

2.4.3 Explainability methods

As mentioned before, it is possible to distinguish the methods that produce explanations by their capability to be applied in any model (model-agnostic) or only to a specific family of models (model-specific). The most significant advantage of model-agnostic interpretation methods over model-specific ones is their flexibility. It is easier to work with model-agnostic explanations since it allows ML developers to develop any machine learning model and to compare different ML models in terms of interpretability.

Model-agnostic methods create explanations by generating feature summaries related to their importance degree and the interaction between them. These interpretation methods can also be distinguished between global and local methods.

An alternative to model-agnostic interpretation methods is implementing only intrinsically interpretable models such as linear and logistic regressions, decision trees and decision rules. Although, it has the disadvantage that predictive performance is lost compared to more complex models and limits the developer to one type of ML model.

Another choice is to use model-specific interpretation methods. However, it also restricts the developer to only one type of ML model and is not easy to exchange to another application.

Example-based explanations can also be considered model-agnostic since they render any machine learning model more interpretable. This method selects specific dataset instances to describe the behavior of ML models.

Even if it is possible to use model-agnostic explanations to make deep learning models more interpretable, more specific interpretation methods are needed to interpret the behavior and predictions of neural networks.

2.4.3.1 Global model-agnostic methods

Global methods characterize the average behavior of a machine learning model. They are appropriate to understand the general mechanisms in the data or debug the model. In this section, some examples of such techniques are described.

Partial Dependence Plot (PDP)

Partial Dependence Plot is a global method that helps to interpret complex models by plotting the impact of specific features or a subset of features on a model's prediction. It shows how a particular set of features influences the average predicted value by marginalizing the remaining features [8, 10].

In mathematical terms, the partial dependence function ($\hat{f}_{\mathrm{S}}$) is defined by equation 2.10 [10].

$$\hat{f}_{\mathrm{S}}(x_{\mathrm{S}}) = E_{\mathrm{X_C}}\big[\hat{f}(x_{\mathrm{S}}, X_{\mathrm{C}})\big] = \int \hat{f}(x_{\mathrm{S}}, X_{\mathrm{C}}) dP(X_{\mathrm{C}}) \tag{2.10}$$

Where the x_{S} are the features for which we want to know the effect on the prediction and X_{C} are the remaining features used in the machine learning model.

In Figure 2.14, there is an example where it is possible to visualize the influence of weather features (temperature, humidity and wind speed) on the predicted number of bikes rented on a given day. Observing the plots, it is possible to conclude that the temperature shows the most significant differences: the hotter, the more bikes are rented until it is too hot, which conducts to a decrease in the rented bikes [10].

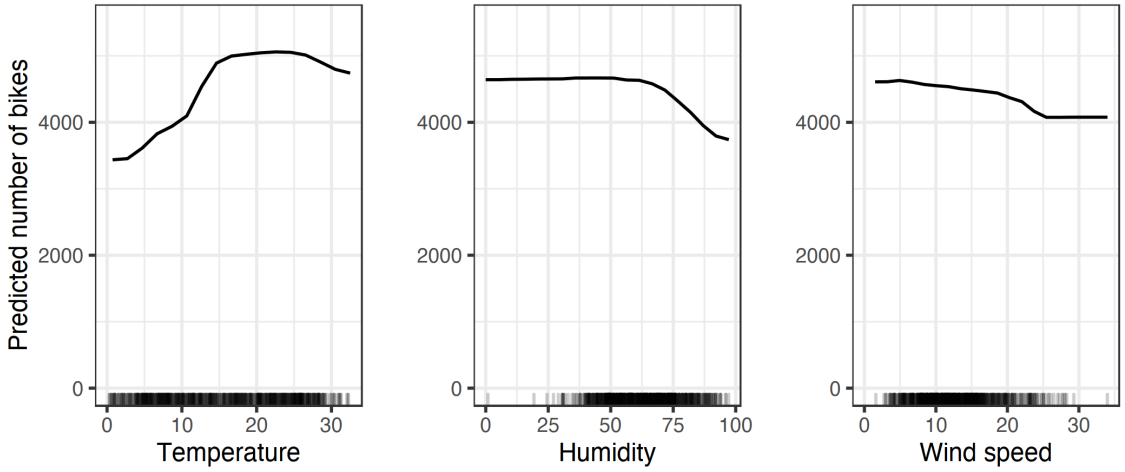

Figure 2.14: PDP for the prediction count model of bicycle renting of weather features (temperature, humidity and wind speed). Source: Molnar et al. [10].

This interpretation method is intuitive and straightforward since it is easy for a layperson to understand the idea of PDP. Furthermore, it can often provide helpful information and is computationally easy to implement. However, PDPs are usually simplistic and do not consider all different feature interactions. It assumes that the features used to compute the partial dependence are not correlated with others. Consequently, PDPs may not provide an accurate approximation to the real relationships between variables. Additionally, this method can only analyse two features simultaneously and shows only the average marginal effects, which might hide heterogeneous effects [8, 10].

A faster and unbiased alternative to PDP is Accumulated Local Effects (ALE) plot. It describes how features influence the model by calculating the average differences in predictions based on the conditional distribution of the features, instead of average predictions based on marginal distributions, as in PDP. Thus, ALE plots try to address the most critical shortcoming of PDPs, the assumption of independence between features. However, the implementation of this method is more intricate [8, 10].

Feature Interaction

The main goal of the feature interaction method is to understand how the relationship between features affects the final prediction. This global interpretation method measures the interaction strength between features using Friedman's H statistics. It evaluates the existence and extent of the interaction between two features or between a feature and the remaining ones.

H-statistics can be interpreted meaningfully and have the advantage of being dimensionless, which allows comparison across features and even models. Additionally, the statistics identify all types of interactions, regardless of their nature, which is a suitable method to use before creating PDP for relationships of interest. However, the procedure is computationally expensive, and there is no standard threshold from which the interaction is strong enough to be considered relevant [10].

Permutation Feature Importance

Permutation feature importance is a global interpretation method that estimates the importance of a feature by computing the growth in the model's prediction error after permuting the feature's values. A feature is considered "important" if a change in its value increases the model error since the model relies on the feature to make predictions. Otherwise, the feature is considered "unimportant". Permutation feature importance is a straightforward method that gives global insights into

the model's behavior, considering all interactions between features. However, this method can be biased by unrealistic data instances [10].

Global Surrogate

The global surrogate method aims to approximate the predictions of a black-box model as accurately as possible using simple interpretable ML models. The global surrogate is considered a model-agnostic method since it does not require any information about the inner workings of the opaque model. It is applied after the model's training phase and only needs access to data and the prediction function.

The main advantage of the surrogate model method is its flexibility since any interpretable machine learning model can be used. This approach is also very intuitive and easy to implement. However, it is still unclear how well the surrogate model should approximate the black-box model to be trusted. It is also important to note that the conclusions are related to the model and not the data since the surrogate model does not have information about the real outcome [10, 47].

Prototypes and Criticisms

A prototype is a representative instance of all the data. Data instances that are inadequately represented by the prototype set are considered criticisms. It can provide insights into complex data distribution together with prototypes, especially for data points not well represented by the latter. Figure 2.15 shows a simulated data distribution and the chosen instances for prototypes and criticisms.

These two concepts can be combined in a single framework by the Maximum Mean Discrepancy (MMD)-critic approach. It employs the MMD statistic to measure the similarity between points and potential prototypes and accurately selects prototypes that maximize the statistic. Additionally, MMD-critic picks criticisms samples applying a regularized witness function score.

Apart from helping to understand the data distribution, MMD-critic can also be used to create an interpretable model or to make a black-box model interpretable.

This technique is easy to implement and works with any data and machine learning model. But it does not consider irrelevant features, and the distinction between prototypes and criticisms depends only on the chosen number of prototypes [10, 50].

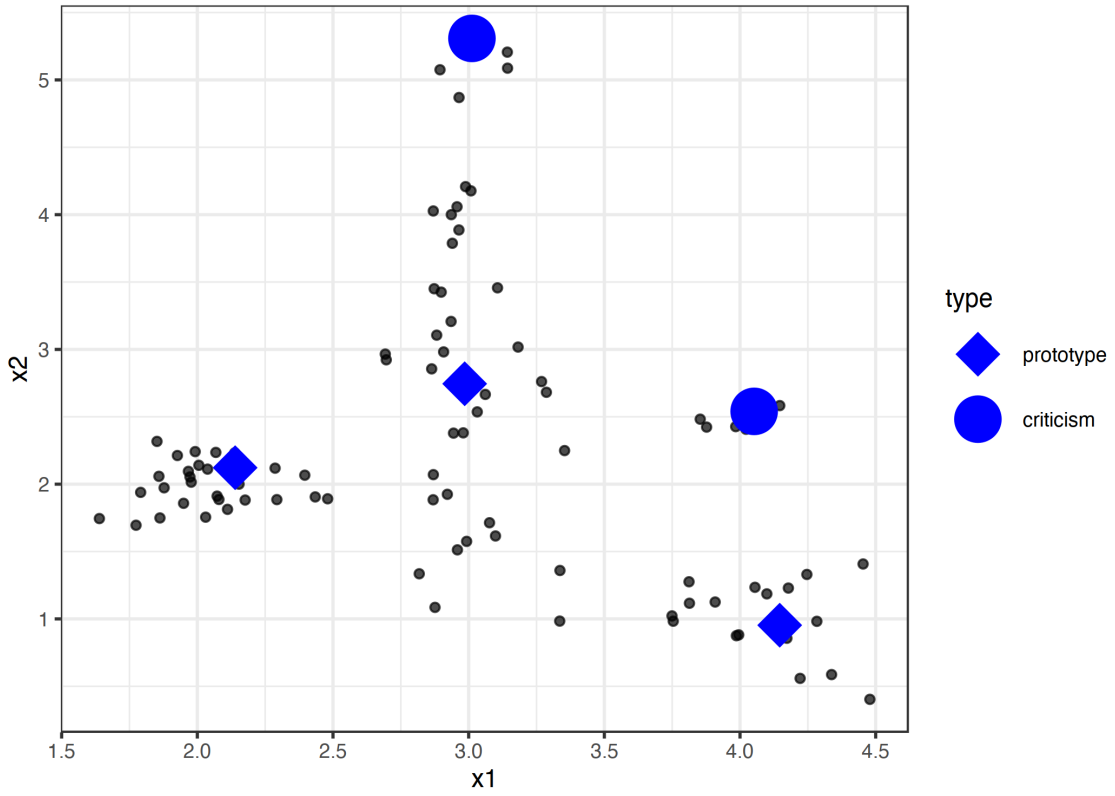

Figure 2.15: Prototypes and criticisms for a data distribution with two features. Source: Molnar et al. [10].

2.4.3.2 Local model-agnostic methods

Local interpretation methods provide explanations for individual predictions. Some examples of such techniques are described in this section.

Individual Conditional Expectation (ICE) plots

The Individual Conditional Expectation (ICE) plot is the equivalent to a PDP for individual data instances, being, therefore, a local method. An ICE plot displays the dependence of the prediction on a feature for each sample, producing one line per instance. As a result, ICE curves can uncover the heterogeneity of the relationships unlikely to PDPs. Although, it can only display one feature at a time in a meaningful way. Furthermore, it might be challenging to distinguish relevant characteristics since the plot can become overcrowded if many ICE curves are drawn [8, 10].

Considering the bicycle rental prediction example, we can conclude, observing Figure 2.16, that ICE curves present the same effects as PDPs. Therefore, the partial dependence plot is a good summary of the relationships between the weather features and the predicted number of rented bikes [10].

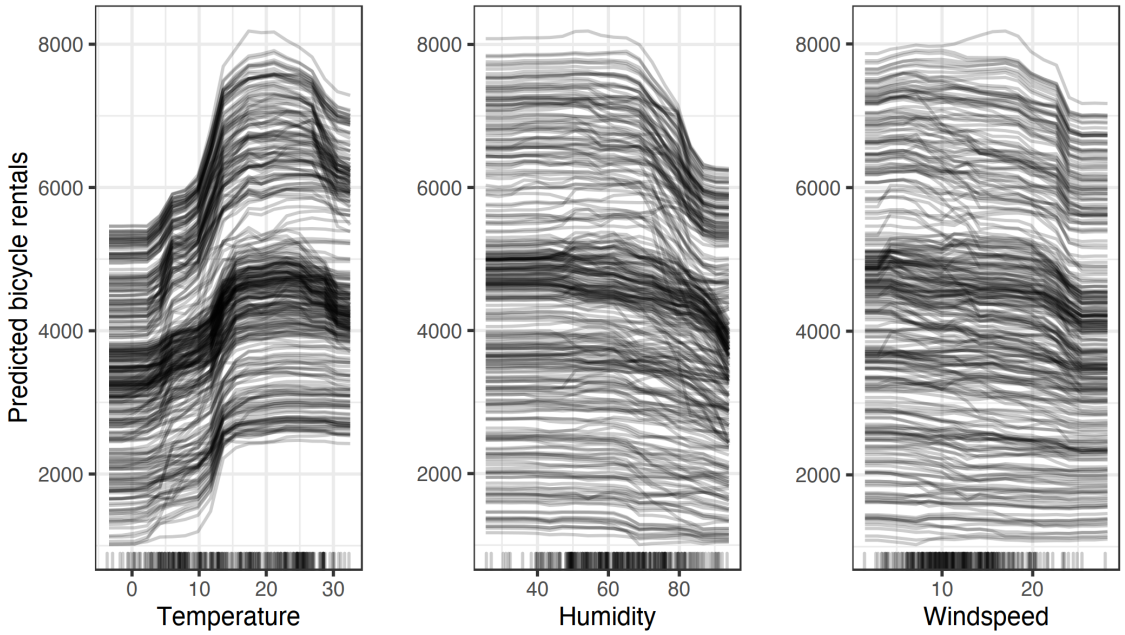

Figure 2.16: ICE plots for the prediction count model of bicycle renting of weather features (temperature, humidity and wind speed). Source: Molnar et al. [10].

Local Surrogate (LIME)

The local surrogate method aims to explain individual predictions of black-box machine learning models by training interpretable models. The idea of Local Interpretable Model-agnostic Explanations (LIME) is quite intuitive. It tests what happens to the predictions when data variations are supplied to the machine learning model. A new dataset is generated, consisting of perturbed samples and the corresponding predictions of the black-box model. Then, LIME fits an interpretable model, weighted based on the nearness of the sampled instances to the instances of interest. The trained model should be a good approximation of the ML model locally but not globally.

LIME is an up-and-coming method since it produces short and straightforward explanations and is easy to use. However, it still has many drawbacks that need to be solved to be safely applied [8, 10, 47].

Scoped Rules (Anchors)

The anchors method explains single predictions of any black-box machine learning model by creating decision rules that sufficiently determine ("anchors") the local prediction. It means that changes to other feature values do not impact the prediction value. Like the LIME method, the anchors method implements a perturbation-based procedure to generate local explanations for predictions of opaque models.

Although, instead of surrogate models, the local explanations are represented as IF-THEN rules, called anchors.

The anchors approach produces explanations easy to interpret, even by laypeople. However, it suffers from a highly configurable and impactful setup, and also, many cases need discretization as otherwise, results will be too specific [8, 10]

Shapley Values

Shapley value is a local interpretation method based on game theory. This method produces explanations by assuming, for each instance, that each feature is a "player" and the prediction is the "payout". The main idea of the Shapley value is for a given feature in the instance to be explained, evaluating its contribution across all possible coalitions (sets) of features. It aims to distribute the payout fairly among them.

This approach requires much computing time since the number of coalitions increases exponentially with the number of features. Furthermore, when variables are correlated, it can include unrealistic data instances.

It should be stressed that the Shapley value is the average contribution of a feature to the model prediction across different coalitions and not the difference in prediction when the feature is removed from the model [10, 47].

Considering the bicycle renting example, Figure 2.17 displays the Shapley values of each feature regarding an instance. For this instance, it is possible to observe that weather and humidity had the most significant negative contribution to the prediction, while temperature had the most critical positive contribution. The sum of Shapley values renders the difference between actual and average prediction (-2108) [10].

Counterfactual Explanations

Counterfactual explanations can be used to explain predictions of individual instances. It describes the slightest possible change to the feature values that alter the prediction to a predefined output. Thinking in counterfactuals requires creating a hypothetical situation that contradicts the observed facts. The goal of this approach is not to reveal the model's inner workings but to identify the factors that can be changed to produce the desired outcome.

This method is relatively easy to implement, and its explanations are easy to understand by humans. The counterfactual method does not require access to the data or the model and works with systems that do not employ machine learning.

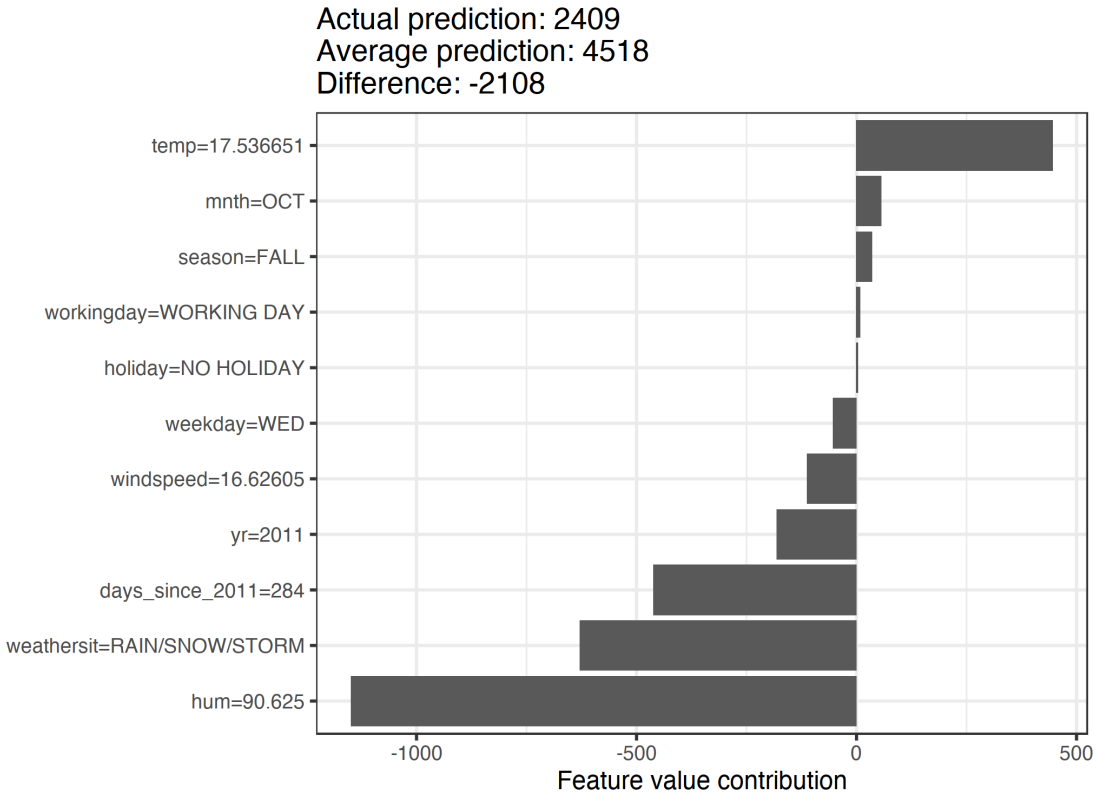

Figure 2.17: Shapley values regarding an instance from the prediction count model of bicycle daily renting. Source: Molnar et al. [10].

However, it may not be appropriate and sufficient in specific scenarios. Furthermore, for each instance, multiple counterfactual explanations can be found [8, 10].

2.4.3.3 Example-based methods

Example-based methods, as mentioned before, explain the behavior of ML models or the underlying data distribution, selecting particular instances of the dataset. Some of the methods presented above are example-based, including counterfactual explanations, prototypes and criticisms. Other methods are K-Nearest Neighbors (KNN) model, an interpretable ML model, and the influential instances method described in this section.

Influential Instances

This technique aims to identify influential training instances that considerably change the parameters or predictions of the model when deleted from the training data, as illustrated by Figure 2.18.

One popular technique is deletion diagnostics, in which individual training in-

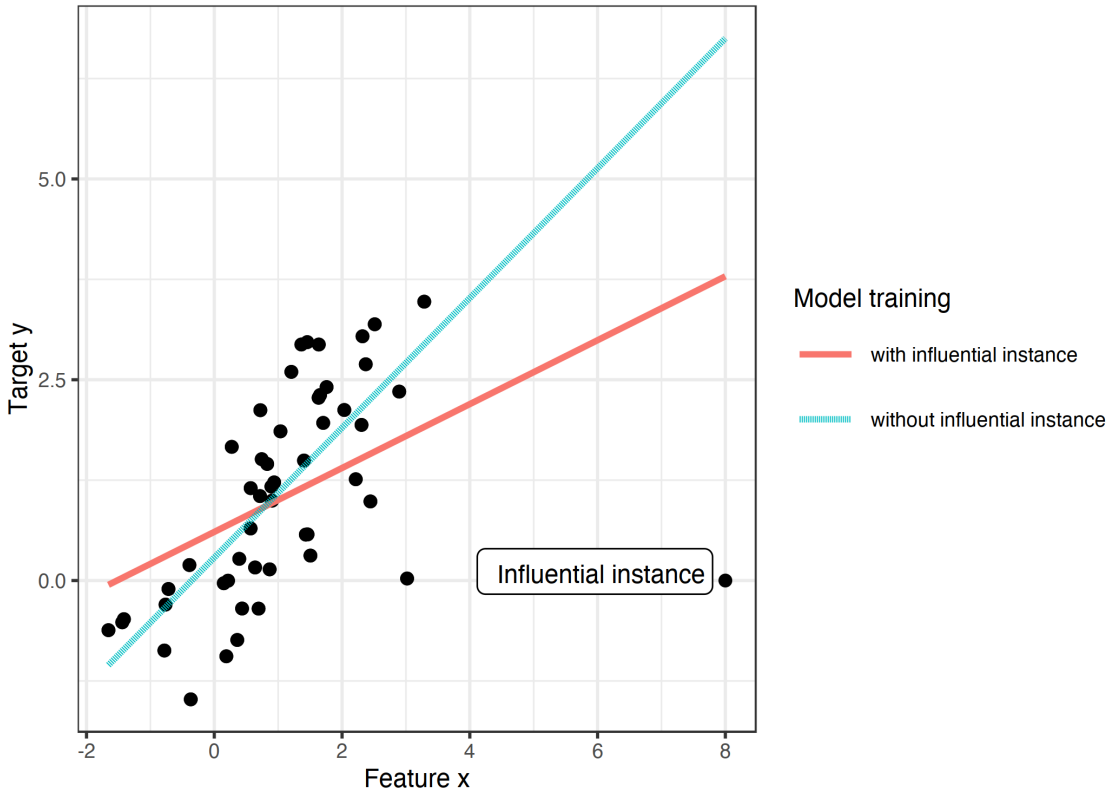

Figure 2.18: Representation of a linear model with one feature. Trained once with the full data and once without the influential feature. Source: Molnar et al. [10].

stances are omitted one at a time, and the model is retrained repeatedly. Then, the parameters or predictions of the original model are compared with those of the retrained model. However, this technique may be problematic regarding computing time since the model needs to be retrained for each training instance. Another possible approach is influence functions, which use robust statistics to approximate how much the model changes when the weight of the sample is increased. This method helps to comprehend the model behavior, debug the model and identify errors in the dataset. However, it requires access to the loss gradient regarding the model parameters, which is only possible for a specific group of ML models [10].

2.4.3.4 Deep Learning interpretation

Although deep learning models are not implemented in this thesis, a brief overview of deep learning interpretation methods is presented in this section. Most studies on interpretability and explainability in the seizure prediction field are conducted using deep learning models, as shown in the state-of-the-art chapter.

Deep learning has acquired popularity over the years since it can outperform state-of-the-art accuracy, frequently exceeding human-level performance. Subse-

quently, many real-world problems in diverse fields are being addressed by deep learning models. However, explaining and interpreting their predictions and behavior is a challenging task. These networks include several layers and weights through non-linear transformations that provide high complexity to their internal mechanisms. As a result, interpretation methods produced particularly for neural networks are needed. In the first place, neural networks learn features and concepts in their hidden layers that universal models may not uncover. Secondly, the gradient can be used to implement interpretation methods that are more efficient than model-agnostic ones, that only consider the model from the outside [10].

The deep learning interpretation methods are usually designed with distinct principles [51]:

- underlining the features on which the deep model mainly relies, with gradients, perturbations, or explainable proxy models;
- analyzing the inside of deep models to understand the logic mechanisms;
- evaluating the contributions of each training data instance for interpreting the inner mechanism.

2.4.4 Grounded Theory (GT)

Explanations are social interactions where beliefs are exchanged. People may have different criteria to identify a proper explanation [10]. As a result, evaluating explanations produced by the presented methods can be challenging, especially on a quantitative level. On that account, the developed decision explanations can be presented to domain experts and evaluated at the human-grounded level. In order to arrive at theoretical explanations, the impressions given by specialists should be rigorously analyzed using tools such as Grounded Theory (GT), the procedure most widely used in scientific research.

Grounded theory is an inductive research methodology designed to provide hypotheses and theoretical explanations about social phenomena for which slight theory has been developed. The theory is grounded in the systematic collection and analysis of data that are mainly, but not exclusively, qualitative. The survey participants' experiences conduct the research, and the results reflect patterns in these experiences. Researchers must avoid preconceived assumptions and adopt an impartial view regarding the topic under debate since GT aims to develop a theory or explain a process, not test or verify an existing hypothesis [52, 53].

The grounded theory employs an interactive approach that implicates cycles of simultaneous data collection and analysis, in which the emerging results of the data

analysis are used to inform subsequent data collection. The process continues until it achieves theoretical saturation, which identifies the point where no new information is obtained in the data collection, as represented in Figure 2.19 [52].

The data is commonly acquired in grounded theory studies by interviews or observational fields. Since the goal of data collection is to obtain an appropriately wide range of perspectives and experiences relevant to the research questions, the data sampling must provide information that would confirm, challenge or expand an arising theory. Therefore, the data sources are selected as the data analysis progress until saturation is achieved. In GT, having a large number of interviews or data is not crucial, but it is essential to have the correct quantity to ensure saturation [52].

The analytical procedure is based on the constant comparison of the data and consists of four stages [54].

In the first stage, recognized incidents are compared against other examples for similarities and differences. Similar incidents are grouped into large categories using a procedure known as open coding. At this point, data is split, labelled and placed into the appropriate categories [54]. The themes are constantly redefined, through systematic comparisons that rename and reorganize the categories, according to the ongoing data collection [52]. While comparing instances and coding, researchers write in memo forms possible patterns identified in and between codes. It helps to organize and formulate emergent theory at progressive levels [52, 53].

In the second stage, connections and properties are attempted to be established within each category. It is achieved by axial coding that groups the initial themes into key elements [54].

In the third stage, similar categories are grouped into highly conceptual themes, and hypotheses are generated until saturation is reached [54].

In the last stage, the writing of the theoretical formulations takes place. The analytical process is considered complete when the theory produces understandable explanations of the social phenomenon under study [52, 54].

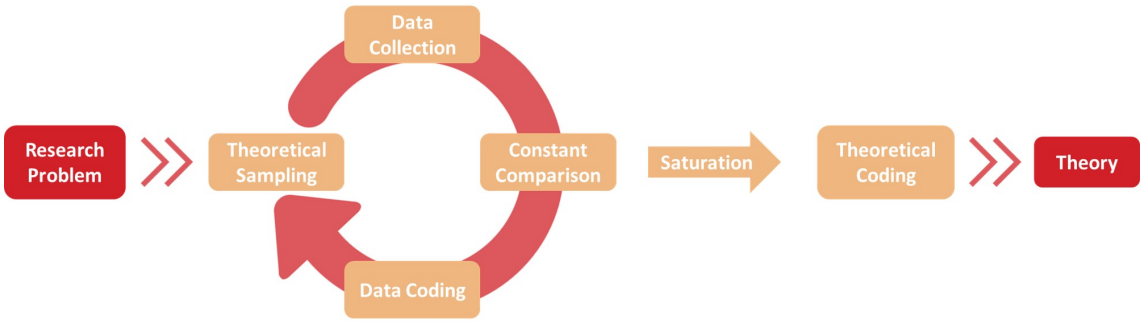

Figure 2.19: Grounded Theory (GT) flow chart.

2.5 Summary

Epilepsy is one of the most common neurological diseases expressed by abnormal brain activity resulting in seizure events. It is characterized by significant heterogeneity concerning types of seizures, epilepsy, and epilepsy syndromes. A seizure can also be categorized by its initial manifestations/symptoms, awareness and epileptic focus localization involving lobes and/or hemispheres.

Patients with DRE, who cannot control seizure activity through medication, are the main focus of seizure prediction studies. This group of patients is often subjected to pre-surgical monitoring for long periods to evaluate their condition, constituting most of the databases used in epileptic seizure prediction studies.

The EEG is a medical tool widely used in detecting and analyzing epileptic seizures since it can measure and record the brain's electrical activity. Two different methods can be used to acquire the signal: scalp EEG and iEEG. While iEEG recordings present a higher signal-to-noise ratio, scalp EEG recordings can capture low-frequency activity more accurately. A typical EEG record from an epileptic patient can be divided into preictal, ictal, postictal and interictal.

The goal of seizure prediction is to correctly anticipate a seizure by detecting the preictal period providing a well-defined occurrence period (SOP) with a predefined horizon (SPH). However, this transitional stage differs between patients and seizure episodes, presenting a significant challenge.

Sensitivity and FPR/h are the gold standard metrics to evaluate the seizure prediction algorithms' performance. Furthermore, statistical validation plays a fundamental role in the evaluation of seizure prediction algorithms as well as in the comparison of distinct prediction models. A proper methodology must involve an

adequate SPH duration, long enough for patients to take preventive actions and assess performance for a range of SOP values, which should not be too long since it may contribute to increasing the patient's anxiety. Finally, robust models should handle the existence of concept drifts and data imbalance.

The deployment of ML models in sectors such as healthcare has led to an increased interest in optimized systems not only at the performance level but also on the level of other relevant criteria, including safety, trustworthiness, fairness, robustness and the right to explanations. The explainability field aims to combat skepticism regarding the clinical application of machine learning models and enhance patient and clinician trust. However, as explainability is still a very recent field of investigation, there is no consensus regarding many aspects, namely, how to organize and evaluate the different explanations and how to define the terms of interpretability and explainability.

3

State of the art

An overview of seizure prediction and explainability studies is presented in this chapter. Section 3.1 describes the general framework underlying EEG seizure prediction algorithms. Section 3.2 presents some recent studies about the explainability of EEG-based models. In Section 3.3, the main ideas of the state of the art are summarized.

3.1 Seizure Prediction

The epileptic seizure prediction field has a brief but rich history of speculation, debate and accomplishments. However, it is still a challenging area that has earned attention from the most diverse disciplines.

The preliminary work on the seizure prediction field dates back to the 1970s. At first, seizure prediction studies relied on statistical methods. With the growth in technology, the capacity to record and develop machine learning models has expanded. Then, with the rapid increase in computing power and computer storage, more complex prediction algorithms have been applied to continuous multi-day EEG recordings, such as deep learning approaches. Therefore, the most diversified research approaches, dataset considerations, evaluation parameters, and implementation techniques have been used since the inception of the seizure prediction field [16, 18, 19].

3.1.1 Framework overview

Most seizure prediction studies follow a series of general steps regarding signal processing, machine learning and post-processing (see Figure 3.1). Despite the existence of this general framework, there is a considerable variety of studies with different approaches regarding used parameters and methods.

Regarding EEG signals, statistical analysis, and ML approaches, each step can be summarized as follows:

- Signal acquisition: data collection to be used in the study.
- Pre-processing: signal preparation for adequate feature extraction - artifacts removal, signal-to-noise ratio improvement and data segmentation through sliding window analysis.
- Feature extraction: extraction of descriptive characteristics from the EEG signal.
- Feature selection: selection of the most relevant features to discriminate each epileptic state.
- Classification: training ML models to identify preictal changes using the selected features.
- Post-processing: regularization methods to smooth the classifier output and give temporal meaning to the consecutive independent predictions.
- Performance evaluation: assessment of the algorithm performance applying the appropriate metrics.

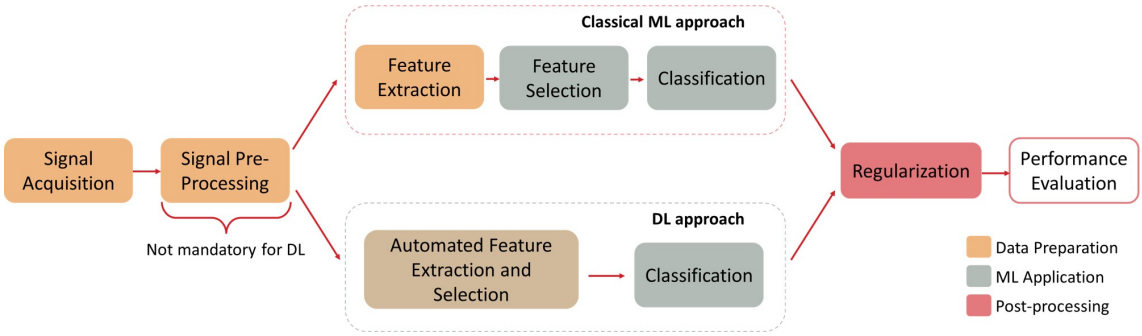

Figure 3.1: General framework of seizure prediction studies. Adapted from: Rasheed et al. [11].

Over the years, various seizure prediction studies have employed classical machine learning approaches. These classification algorithms rely on handcrafted features obtained from traditional signal processing methods. However, the feature extraction process and all the inherent steps require extensive computational time and can discard information from the signal which may be relevant for prediction. Therefore, it is a challenging task to automatically extract informative characteristics from the raw signal concerning the final goal of the study.

On the other hand, deep learning algorithms can automatically learn more distinct and robust features than traditional methods. It is an evolved ML technology capable of rigorously learning patterns from extensive data collections. Furthermore, the capability of deep learning models to produce more accurate results has influenced the researchers to tackle the seizure prediction problem by employing DL

techniques.

However, once these models present more parameters to be trained, the risk of overfitting is higher, requiring larger datasets. Nevertheless, less data is available to train such models with the evolution of patient-specific algorithms. Furthermore, its lack of interpretability has raised some skepticism regarding its clinical applicability [11, 55].

Therefore, as depicted in Figure 3.1, the leading adaptations introduced by DL models in the seizure prediction framework are related to feature extraction, feature selection and classification steps. These stages can be pooled together since deep learning techniques can handle raw data.

3.1.2 Signal acquisition

Signal acquisition and data selection are the first steps in seizure prediction studies. The database used for research plays a significant role in the performance of the designed approach.

Various hospitals and research centres have published different databases over the years. The main differences between databases are related to the EEG recording mechanisms used, the number of subjects used, and the number of channels used. Regarding these aspects, Table 3.1 presents an overview of the EEG databases used in seizure prediction studies over the past ten years.

Databases

The most widely used databases in the EEG seizure prediction studies are the European Database on Epilepsy (EPILEPSIAE) [56–63, 66, 75, 80, 85], the Children's Hospital Boston database (CHB-MIT) [64, 67, 69–71, 74–77, 79, 81–84] and the Freiburg database [65, 72, 73, 75, 78, 82].

The EPILEPSIAE database is the most significant accessible from pre-surgical monitoring, comprising EEG recordings lasting 165 hours (on average) from 275 DRE patients. Additionally, it includes standardized annotation and extensive metadata.

On the other hand, the NeuroVista database created by Cook et al. [24] is the largest one regarding recording duration per patient. It comprises data from 15 patients who were followed for up to two years outside monitoring units. Therefore, this database is a good representation of real-life data and should benefit from concept drift and promote clinical translation.

Freiburg Hospital's database is one of the substantial databases which contains

Table 3.1: Overview of the signal acquisition aspects underlying EEG seizure prediction studies over the past 10 years.

Study	Database	Number of Patients	Electrodes	EEG Type
Teixeira et al. [56] (2012)	EPILEPSIAE	10	6 covering the scalp, 3 focal and 3 afocal	Scalp EEG
Bandarabadi et al. [57] (2012)	EPILEPSIAE	12	3 focal and 3 afocal	Scalp EEG, iEEG
Cook et al. [24] (2013)	NeuroVista	15	16	iEEG
Rasekhi et al.[58] (2013)	EPILEPSIAE	10	3 focal and 3 afocal	Scalp EEG, iEEG
Rabbi et al. [59] (2013)	EPILEPSIAE	1	2 focal	iEEG
Alvarado-Rojas et al. [60] (2014)	EPILEPSIAE	53	-	iEEG
Teixeira et al. [61] (2014)	EPILEPSIAE	278	6 random, 6 covering the scalp, 3 focal and 3 afocal	Scalp EEG, iEEG
Rasekhi et al. [62] (2015)	EPILEPSIAE	10	3 focal and 3 afocal	Scalp EEG, iEEG
Bandarabadi et al. [63] (2015)	EPILEPSIAE	24	3 focal and 3 afocal	Scalp EEG, iEEG
Usman et al. [64] (2017)	CHB-MIT	24	23	Scalp EEG
Aarabi et al. [65] (2017)	Freiburg	10	3 focal and 3 afocal	iEEG
Dieito et al. [66] (2017)	EPILEPSIAE	216	6 random, 6 covering the scalp, 3 focal and 3 afocal	Scalp EEG, iEEG
Khan et al. [67] (2017)	MSSM, CHB-MIT	28, 22	22	Scalp EEG
Kiral-Kornek et al.[68] (2018)	NeuroVista	10	16	iEEG
Tsiouris et al. [69] (2018)	CHB-MIT	24	18	Scalp EEG
Usman et al. [70] (2018)	CHB-MIT	24	23	Scalp EEG
Kitano et al. [71] (2018)	CHB-MIT	9	23	Scalp EEG
Yuan et al. [72] (2018)	Freiburg	21	3 focal	iEEG
Yang et al. [73] (2018)	Freiburg	19	6	iEEG
Daoud et al. [74] (2019)	CHB-MIT	8	23	Scalp EEG
Truong et al. [75] (2019)	CHB-MIT, Freiburg, EPILEPSIAE	13, 13, 30	16, 3 focal and 3 afocal, 19	Scalp EEG, iEEG, Scalp EEG
Zhang at al. [76] (2019)	CHB-MIT	23	18	Scalp EEG
Gabara et al. [77] (2020)	CHB-MIT	6	23	Scalp EEG
Stojanović et al. [78] (2020)	Freiburg, Epylepsyecosystem	5, 3	31-122, 16	iEEG
Tamanna et al. [79] (2021)	CHB-MIT	10	23	Scalp EEG
Pinto et al. [80] (2021)	EPILEPSIAE	19	19	Scalp EEG
Usman et al. [81] (2021)	CHB-MIT	22	23	Scalp EEG
Peng et al. [82] (2022)	CHB-MIT, Freiburg	16, 20	23, 3 focal and 3 afocal	Scalp EEG, iEEG
Singh et al. [83] (2022)	CHB-MIT	24	23	Scalp EEG
Liang et al. [84] (2022)	CHB-MIT, Kaggle (American Epilepsy Society)	12, 5 dogs and 2 patients	23, 16 (dogs), 15 depth (pat1), 24 subdural (pat2)	Scalp EEG, iEEG
Pinto et al. [85] (2022)	EPILEPSIAE	93	19	Scalp EEG

intracranial EEG recordings of 21 subjects with around 88 seizures. Still, lately, it has been integrated into the EPILEPSIAE database to provide larger datasets. The database from the Center of Epilepsy at Children's Hospital, Boston, includes scalp EEG recordings from 23 patients divided into seizure and non-seizure recordings. Additionally, some studies [84] have incorporated epileptic animals such as dogs included in the Kaggle American Epilepsy Society database.

More recently, new databases have started appearing with long-term data and acquisition methods that are more comfortable for patients. For instance, Attia et al. [86] used ultra-long-term subcutaneous EEG data in their study. New subcutaneous EEG recording systems have shown promise in enabling continuous monitoring with a modest burden to the patient. On the other hand, Nasseri et al. [87] used ultra-long-term recordings from a noninvasive wrist-worn multimodal sensor. Accelerometry (ACC), blood volume pulse (BVP), electrodermal activity (EDA), temperature (TEMP) and heart rate (HR) signals were recorded by the wearable device and used in a seizure forecasting system.

Recording type and channels selection

Regarding the type of EEG recording, both intracranial and scalp EEG have been commonly used. Indeed, some studies [56–58, 61–63, 66, 75, 82, 84] employed the two recording types and analyzed the impact on the results. However, non-significant discrepancies were reported.

Distinct strategies have been used regarding electrode selection. While some researchers work with all available electrodes, others select a given number of electrodes from specific or random locations. Choosing all available electrodes can be intuitive since it contains all the information. However, the large number of electrodes may lead to high computational costs and patient discomfort.

The location selection is determined by different assumptions concerning the seizure generation process. The random electrode choice [61, 66] is based on the presumption that seizure generation can be captured in any brain region. Otherwise, choosing electrodes only from the focal area [59, 72] presumes that the most discriminative traces are given by those placed near the seizure focus. When the electrodes are chosen from the focal and far from the focal regions [56–58, 61–63, 65, 66, 75, 82], it is assumed that the preictal stage is adequately represented by relating the most discriminative traces with information of the general brain state provided by afocal electrodes. Moreover, to capture generalized brain electrical activity, electrodes can be chosen to maximize coverage of the scalp [56, 61, 66].

3.1.3 Signal pre-processing

Pre-processing is a crucial procedure in raw EEG signals. It aims to remove noise and artifacts in order to reduce their influence in feature extraction and consequently improve the classification performance of the designed algorithm. Numerous techniques have been employed to minimize the impact of unwanted artifacts and noise and prepare the signal for feature extraction [11, 88].

Typical strategies include artifact removal and filtering, data segmentation performed by sliding window analysis, and definition of Seizure Occurrence Period (SOP) and Seizure Prediction Horizon (SPH) duration. Although these two periods can be defined right before the model training, they must be selected during the initial steps according to the application system, as explained in Section 2.3.1, and not based on the model's performance. Therefore, their definition was incorporated in the signal pre-processing phase.

Table 3.2 presents an overview of the pre-processing procedures adopted in seizure prediction studies over the past ten years.

Filtering and artifact removal

This step involves the removal of environmental noise, high-frequency content, and artifacts such as eye movements, muscle activity and cardiac signals.

The power-line interference is commonly removed with a notch filter of 50 Hz [56, 58, 61–63, 65, 80, 85] or 60 Hz [59]. Band-pass filters are also frequently implemented [24, 59, 60, 65, 66, 75, 76, 80, 82–84], although their cut-off frequencies depend on the bands of interest that differ among studies. Typically, low-frequency components below 0.5 Hz are removed since they are considered breathing artifacts. However, high-frequency components which involve environmental noise are harder to delimit. The most frequent strategy is filtering with Infinite Impulse Response (IIR) and Finite Impulse Response (FIR) digital filters.

Despite artifact removal being considered an essential step in classical ML approaches, it can be viewed as risky since helpful information from the signal can be discarded along with the artifacts [55].

Data segmentation

The EEG recordings should be segmented into smaller windows to extract relevant features from the signal. Most importantly, in a real scenario, the data segmentation allows online analysis, permitting decisions in short periods.

These windows are supposed to comprise similar characteristics significant to

Table 3.2: Overview of the pre-processing procedures adopted in the EEG seizure prediction studies over the past 10 years.

Study	Denoising, Filtering, Artifact Removal	Sliding Window Length (seconds)	Pre-ictal period (minutes)	SPH
Teixeira et al. [56] (2012)	50 Hz Notch	5	10, 20, 30, 40	-
Bandarabadi et al. [57] (2012)	-	5	10, 20, 30, 40	-
Cook et al. [24] (2013)	Octave-wide digital and Notch filters 8 Hz - 128 Hz	5	-	-
Rasekhi et al. [58] (2013)	50 Hz Notch	5	10, 20, 30, 40	-
Rabbi et al. [59] (2013)	0.5 - 100 Hz Butterworth, 60Hz Notch	10 (50% overlap)	15, 30, 45	-
Alvarado-Rojas et al. [60] (2014)	8th-order Butterworth filter in bands of interest from 0.5Hz to 140 Hz, Hilbert transform	60	10, 30, 60	-
Teixeira et al. [61] (2014)	50 Hz Notch	5	10, 20, 30, 40	-
Rasekhi et al. [62] (2015)	50 Hz Notch	5	10, 20, 30, 40	-
Bandarabadi et al. [63] (2015)	50 Hz Notch	5	10, 20, 30, 40	-
Usman et al. [64] (2017)	-	8	-	-
Aarabi et al. [65] (2017)	0.5 - 100 Hz Butterworth, 50 Hz Notch	10	30, 50	10 s
Direito et al. [66] (2017)	48 - 52 Hz Butterworth	5	10, 20, 30, 40	10 s
Khan et al. [67] (2017)	128 Hz low-pass filter	1	10	-
Kiral-Kornek et al. [68] (2018)	-	-	15	-
Tsiouris et al. [69] (2018)	-	5	15, 30, 60, 120	-
Usman et al. [70] (2018)	CSP, EMD	1	-	-
Kitano et al. [71] (2018)	-	4	10	-
Yuan et al. [72] (2018)	-	4	30, 50	10 s
Yang et al. [73] (2018)	Notch filter	5	-	-
Daoud et al. [74] (2019)	-	5	60	-
Truong et al. [75] (2019)	Band-pass filters: 47 - 53 Hz and 97 - 103 Hz or 57 - 63 Hz and 117 - 123 Hz, DC removed	28	30	5 min
Zhang at al. [76] (2019)	5 - 50Hz Butterworth	5	30	-
Gabara et al. [77] (2020)	-	4		-
Stojanović et al. [78] (2020)	Parks-McClellan optimal equiripple FIR filter, Butterworth IIR filter	20 (50% overlap)	5, 60	30 s, 5 min
Tamanna et al. [79] (2021)	-		30 (Preictal/ictal)	-
Pinto et al. [80] (2021)	50 Hz Notch, 0.1 - 20 Hz bandpass	5	40, 50, 60	10 min
Usman et al. [81] (2021)	EMD	29	-	-
Peng et al. [82] (2022)	Band-pass filters: (CHB) 57 - 63 Hz and 117 - 123 Hz, (FSP) 47 - 53 Hz and 97 - 103 Hz	5	30	0
Singh et al. [83] (2022)	0.1 - 127 Hz Butterworth	5, 10, 15, 20, 25, 30	30	5 min
Liang et al. [84] (2022)	Band-pass filters: 57 - 63 Hz and 117 - 123 Hz, DC removal	30 (S samples overlap), 30	-	-
Pinto et al. [85] (2022)	0.5 Hz high-pass filter, 50 Hz Notch filter	5	30-75	10 min

EEG analysis. Therefore, the window size and the percentage of overlap between consecutive windows should be defined considering the compromise between the capacity to capture specific patterns and stationary assumptions [17].

As outlined in Table 3.2, the window length varies between 5 and 30 seconds, with or without overlap, commonly with an overlap percentage of 50%. Even so, it is evident that 5-second windows without overlaps are more frequently adopted [24, 56–58, 61–63, 66, 69, 73, 74, 76, 80, 82, 83, 85].

Pre-ictal period, SOP and SPH duration

No standard or optimal value has yet been defined for the duration of the preictal period. Therefore, the variety of values used in the literature is immense. As shown in Table 3.2, various preictal times have been adopted, ranging from 5 minutes to 2 hours. Some studies adopted a fixed preictal period [67, 68, 71, 74–76, 82, 83], while others considered several values for the preictal period [56–63, 65, 66, 69, 72, 80, 85]. The heterogeneity regarding the preictal duration makes it difficult to compare and evaluate different seizure prediction approaches.

Regarding SPH duration, there is also no standard or optimal value since it may vary depending on the final application. In the literature, SPH ranges from 10 seconds to 10 minutes, as shown in Table 3.2. In most studies, SPH is not mentioned, and it is assumed that this period is dismissed. However, it represents an unrealistic scenario as real-life applications require some time for intervention to be initiated.

The adopted preictal period involves the duration of the SOP and SPH. Hence, when the SPH period is not considered, the SOP time corresponds to the preictal period.

3.1.4 Feature Extraction

The feature extraction step aims to capture the most appropriate discriminant measures that characterize the EEG signal. This step is remarkably heterogeneous due to the immense diversity of features suggested and adopted by researchers. However, no optimal feature type has been identified.

Feature extraction is a crucial factor in the algorithm performance and the prediction model's interpretability and explainability. Consequently, this procedure plays a fundamental role in the trustworthiness and acceptance of seizure prediction algorithms since features with some biological/medical meaning are easily understood by clinical experts.

A feature can be considered univariate or multivariate based on the number

of EEG channels. While univariate features are derived from one single electrode, multivariate features involve information from two or more electrodes. Additionally, the extracted characteristics can be described as linear or non-linear according to the linearity of the captured signal dynamics. Considering both classifications, features can be grouped into four classes: univariate linear, univariate non-linear, multivariate linear and multivariate non-linear, as illustrated in Figure 3.2.

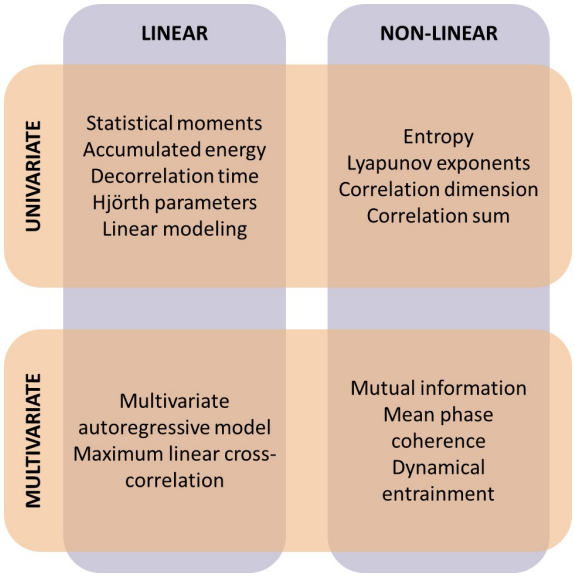

Figure 3.2: EEG features most commonly used in seizure prediction studies, categorized in terms of number of channels and linearity.

Table 3.3 presents an overview of the features type adopted by seizure prediction studies over the past ten years.

It is possible to observe that univariate linear features are widely used in the literature [24, 56–58, 61, 62, 64, 66, 70–72, 77, 79, 80, 83, 85]. The predominance of this feature type may be due to its simplicity, as it is computationally lighter and easier to interpret by clinicians.

It is also important to note that various studies which adopt deep learning models have chosen to apply automatic feature engineering techniques using the raw data as input [69, 74]. However, some still perform traditional feature engineering [71, 83] or transformations to frequency or time-frequency domain [67, 68, 75, 76, 81, 82, 84] using techniques, such as Fourier transform or wavelet decomposition, which are used as input for the classification model. Comparatively, handcrafted features are more interpretable than automatically extracted ones.

Table 3.3: Overview of the features type used in EEG seizure prediction studies over the past ten years.

Study	Univariate		Multivariate		Other
	Linear	Non-Linear	Linear	Non-Linear	
Teixeira et al. [56] (2012)	X	X			
Bandarabadi et al. [57] (2012)	X		X		
Cook et al. [24] (2013)	X	X	X		
Rasekhi et al. [58] (2013)	X				
Rabbi et al. [59] (2013)		X		X	
Alvarado-Rojas et al. [60] (2014)		X			
Teixeira et al. [61] (2014)	X				
Rasekhi et al. [62] (2015)	X		X		
Bandarabadi et al. [63] (2015)			X		
Usman et al. [64] (2017)	X				
Aarabi et al. [65] (2017)		X		X	
Direito et al. [66] (2017)	X				
Khan et al. [67] (2017)					Wavelet transformed EEG
Kiral-Kornek et al. [68] (2018)					Spectogram transformation
Tsiouris et al. [69] (2018)					Raw data
Usman et al. [70] (2018)	X				
Kitano et al. [71] (2018)	X				
Yuan et al. [72] (2018)	X				
Yang et al. [73] (2018)		X			
Daoud et al. [74] (2019)					Raw data
Truong et al. [75] (2019)					From raw data to STFT
Zhang at al. [76] (2019)					From raw data to CSP
Gabara et al. [77] (2020)	X				
Stojanović et al. [78] (2020)					Nonnegative Matrix Factorization
Tamanna et al. [79] (2021)	X	X			
Pinto et al. [80] (2021)	X				
Usman et al. [81] (2021)					From raw data to STFT
Peng et al. [82] (2022)					From raw data to STFT
Singh et al. [83] (2022)	X				
Liang et al. [84] (2022)					(CHB-MIT) From raw data to STFT (Kaggle) From raw data to frequency-domain and time domain by FFT and PCA
Pinto et al. [85] (2022)	X				

3.1.5 Feature Selection

In an attempt to encompass brain dynamics during the transition between states, prediction algorithms usually combine numerous features, which may result in high dimensional feature space. Therefore, it is fundamental to select the most discriminative features that will allow the detection of preictal states. It is a crucial step since irrelevant or redundant features can lead to model overfitting or degrade the classifier performance [17].

Diverse feature selection techniques have been applied in seizure prediction studies, such as maximum Difference Amplitude Distribution of histogram (mDAD) [57, 63], minimum Redundance Maximum Relevance (mRMR) [57, 62, 63], minimum normalized difference of percentiles, ReliefF and Principal Component Analysis (PCA) [17]. Genetic Algorithms (GAs) can also be used in this step [80, 85] and attempt to reproduce the principles of biological evolution: from a random initial population, the strongest one will recombine to survive and adapt to the external environment. When using DL techniques, some authors employ convolution layers [75, 81] or autoencoders [74, 82] to reduce the dimensionality of the feature space.

3.1.6 Classification

Based on the extracted and selected features, a classification algorithm is used to train a model capable of distinguishing the preictal from the interictal period. After the training phase, the model must be employed to unseen data.

Various algorithms have been used in the literature, from simpler to more complex. While earlier studies adopted thresholding [60, 64, 65] and classical machine learning techniques, later studies started to introduce deep learning approaches [56, 67–69, 74–76, 81–84]. As outlined in Table 3.4, the most commonly used classification models are Support Vector Machines (SVM) [56–58, 61–64, 66, 73, 77–79, 82] and Convolutional Neural Networks (CNNs) [67, 68, 74–76, 81, 83, 84]. The LSTM models are also used in a few studies [69, 74, 81].

Partitioning

To evaluate the actual performance of a prediction algorithm is compulsory to test it on out-of-sample data, which were not used in the training phase. Furthermore, a test should not be associated with the same events (seizures) that the training data did. Therefore, partitioning methods are required to divide the data into training, and testing sets [18].

Table 3.4: Overview of the classification, regularization and performance evaluation characteristics used in EEG seizure prediction studies over the past ten years.

Study	Partition Methods	Classifier	Regularlization	Performance	Statistical Validation
Teixeira et al. [56] (2012)	Training: first 3 seizures Testing: remaining seizures	SVM	Firing Power, Kalman Filter	SS(FP)=77% SS(KF)=84% FPR/h(FP)=1.51 FPR/h(KF)=0.20	-
Bandarabadi et al. [57] (2012)	Training: 3 seizures Testing: the remaining	SVM	Firing Power	SS=76.09% FPR/h=0.15	-
Cook et al. [24] (2013)	Training: first 4 months Testing: remaining duration	KNN, Decision Tree	Smoothing	SS=61%	Time-matched predictor
Rasekhi et al. [58] (2013)	Training: first 3 seizures Testing: remaining seizures	SVM	Firing Power	SS=73.9% FPR/h=0.15	-
Rabbi et al. [59] (2013)	Training: 2 seizures Testing: the remaining	ANFIS	-	SS=80.0% FPR/h=0.46	-
Alvarado-Rojas et al. [60] (2014)	Training: first 2-4 seizures Testing: the remaining	Thresholding	Kalman Filter	SS=46.55% FPR/h=0.94	Random Predictor
Teixeira et al. [61] (2014)	Training: first 3 seizures Testing: remaining seizures	ANN, SVM	Firing Power	SS=70.61% FPR/h=0.34	-
Rasekhi et al. [62] (2015)	Training: first 3 seizures Testing: remaining seizures	SVM	Firing Power	SS=60.90% FPR/h=0.11	Random Predictor
Bandarabadi et al. [63] (2015)	Training: first 3 seizures Testing: remaining seizures	SVM	Firing Power	SS=75.6% FPR/h=0.10	Random Predictor
Usman et al. [64] (2017)	10-fold cross-validation	SVM	-	SS=92.23%	-
Aarabi et al. [65] (2017)	Testing: 1 seizures Training: the remaining	Thresholding	-	SS=89.80% FPR/h=0.12	Random Predictor
Direito et al. [66] (2017)	Training: 2-3 seizures Testing: the remaining	SVM	Firing Power	SS=38.47% FPR/h=0.20	Random Predictor
Khan et al. [67] (2017)	Testing: 1-3 seizures Training: the remaining	CNN	-	SS=87.80% FPR/h=0.142	Random Predictor
Kiral-Kornek et al. [68] (2018)	Training: first 2 months Testing: remaining duration	CNN	-	SS=69.0% TiW=27.0%	Random Predictor
Tsiouris et al. [69] (2018)	10-fold cross-validation	LSTM	-	SS=99.0% FPR/h=0.02	-
Usman et al. [70] (2018)	-	Thresholding	-	-	-
Kitano et al. [71] (2018)	-	SOM	-	SS=98%	-
Yuan et al. [72] (2018)	Training: 1-2 seizures Testing: the remaining	BLDA	Moving average filter, thresholding	SS(SOP:30)=85.11% SS(SOP:50)=93.62% FPR/h=0.08	-
Yang et al. [73] (2018)	-	SVM	Two-step firing power	SS=94.0% FPR/h=0.111	-
Daoud et al. [74] (2019)	Leave-One-Out seizures	CNN, Bi-LSTM	-	SS=99.72% FPR/h=0.004	-
Truong et al. [75] (2019)	Leave-One-Out seizures	GAN, CNN, NN	-	AUC(CHB)=77.68% AUC(FSP)=75.47% AUC(EPI)=65.05%	Random Predictor
Zhang at al. [76] (2019)	Leave-One-Out seizures	CNN	Kalman Filter	SS=92.2% FPR/h=0.12	-
Gabara et al. [77] (2020)	Training: 70% of the data Testing: 30% of the data	SVM	-	SS=95.7% ACC=96.2%	-
Stojanović et al. [78] (2020)	Training: 70% of the data Testing: 30% of the data	SVM	-	SS=95.2% SP=99.4% SS=69.0% SP=78.67%	-
Tamanna et al. [79] (2021)	Training: 80% of the data Testing: 20% of the data	SVM	K-of-n method	ACC=96.38% FPR/h=0.19	-
Pinto et al. [80] (2021)	Training: first 60% seizures Testing: the remaining	LogReg	Firing Power	SS=37.0% FPR/h=0.19	Surrogate Predictor
Usman et al. [81] (2021)	k-fold cross validation	CNN+LSTM	-	SS=93.0% SP=92.5%	-
Peng et al.[82] (2022)	Leave-One-Out seizures	MMD-AAE + SVM	-	SS(CHB)=73% FPR/h(GHB)=0.24 SS(FSP)=76% FPR/h(FSP)=0.19	-
Singh et al.[83] (2022)	Training: 90% of the data Testing: 10% of the data	CNN	-	SS=98.0% SP=96.6%	-
Liang et al. [84] (2022)	(CHB-MIT) Leave-One-Out seizures	CNN	(CHB-MIT) K-of-n method	SS(CHB)=88.3% FPR/h(CHB)=0.04 AUC(Kaggle)=0.86	-
Pinto et al. [85] (2022)	Training: first 3 seizures Testing: remaining seizures	LogReg	Firing Power	SS=16.0% FPR/h=0.21	Surrogate Predictor

No standard guidelines have been defined regarding partitioning methods. As seen in Table 3.4, there is considerable heterogeneity between studies.

The majority of the studies present in Table 3.4 consider that the seizure generation mechanism is different between patients. Thus, patient-specific methods are commonly used in the literature, in which a specific model is trained and tested within each patient's data. Additionally, while some studies don't consider the existence of concept drift, neglecting the order in which seizures occur, others assume there is a time dependence by using earlier seizures to train and later ones to test the models [24, 56, 58, 60–63, 68, 80, 85].

Moreover, class imbalance also constitutes a severe issue in seizure prediction studies. While some authors have handled this problem in the training phase by undersampling (discarding some interictal samples) [58, 61, 63, 66], others have addressed this by artificially generating new preictal samples [75].

Support Vector Machines (SVM)

Support Vector Machines (SVM) are supervised machine-learning methods widely used in seizure prediction studies. It is characterized by a good generalization capability and a few optimized parameters. The model aims to identify an optimal separating hyperplane that maximizes the distance between the closest training points from distinct classes. When classes are not linearly separable, SVMs can produce non-linear decision boundaries by using non-linear kernel functions such as the radial basis function (RBF) [17, 66].

Convolutional Neural Network (CNN)

A convolutional neural network is a deep learning model capable of building high-level representations and automatically learning key features directly from the raw data. This method is designed to handle data presented as multiple arrays, such as images and time-series data. Concerning seizure prediction studies, CNNs can capture short-term temporal dependencies from the EEG signal and automatically extract relevant features.

Regarding the architecture, this network generally comprises several convolutional layers capable of producing feature maps via filtering operations with kernels. Then pooling layers can learn features from these feature maps. Following these layers, it is possible to implement classification layers [67].

Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) network is considered an evolution over the Recurrent Neural Networks (RNNs), capable of handling long-term dependencies. The innovative part of LSTM networks compared to traditional RNNs is the incorporation of "gates" to control more rigorously what information needs to be kept in their memory and what must be discarded. By including the three gates (i.e. input gate, forget gate and output gate), the LSTM network can improve the adjustment to extensive sequences of data series compared to other deep learning techniques.

Regarding the seizure prediction problem, the LSTM networks bear the advantage of learning temporal characteristics of the brain activity during different states while maintaining long-time dependencies [69].

Occasionally, CNN and LSTM networks are used together. In these cases, CNN is used to process and extract the features, which the LSTM then employs for classification, leading to their temporality [81].

3.1.7 Regularization

Considering the output of the classifiers as a good alarm generator may be unrealistic since it is unlikely that it will classify all samples correctly, and the model can be susceptible to noise contained in the data. Furthermore, the data is handled as independent windows, which do not correspond to reality. Therefore, an action is also needed to give meaning to the temporality of the algorithms' decisions.

As a result, after the classification, a regularization step should be conducted to reduce the number of false alarms and improve the specificity of the classifier. The regularization functions smooth the classifier's output taking into account its temporal dynamics. As shown in Table 3.4, the most commonly adopted methods are the Kalman filter [56, 76] and firing power [56–58, 60–63, 66, 73, 80, 85].

The firing power method quantifies the relative number of samples classified as preictal using a sliding window of equal length to the pre-ictal period. If this measure exceeds a normalized threshold, an alarm is raised. Although many studies have used firing power, no optimal threshold has been identified [17].

The Kalman filter idea is based on the state estimation of a linear dynamic system. It is a uni-modal, recursive estimator since it only requires the state from the previous time step and current measurement to predict the current state. An alarm is generated when the filter output crosses a given threshold [56].

Teixeira et al. [56] compared both regularization methods. They reported that the firing power measure was more conservative in raising alarms since it presents

a more extended memory of classification dynamics. Though the Kalman filter still generates more false alarms, its sensitivity is relatively better.

3.1.8 Performance evaluation

The final step is to assess the proposed methodology according to a given set of metrics. In seizure prediction studies, the standard performance metrics are sensitivity and False Positive Rate per Hour (FPR/h), along with statistical validation, as discussed in Section 2.3.2. The performance evaluation must be reported concerning the testing phase. However, as outlined in Table 3.4, other metrics have also been used, such as Area Under the Curve (AUC), sample sensitivity and specificity, and time under warning. Furthermore, the majority of the presented studies lack statistical validation.

It is possible to observe there is significant heterogeneity of results, with FPR/h values ranging from 0.004 to 1.51. It is worth noting that various studies obtained FPR/h values below 0.15, which is considered reasonable in the pre-surgical monitoring context.

It is important to note that a proper comparison between studies is difficult due to the distinct methods used and decisions made throughout the entire pipeline.

3.2 Explainability

Despite being an emerging field of study, explainability is not, in fact, a novel area and has been widely used over the years. Due to safety reasons and skepticism, some authors have tried to employ simple models with few but discriminant features to produce human-understandable results. However, no definition or concept had been attributed to this process until recently.

Throughout the last decades, the appearance of more complex algorithms, the deployment of these methodologies in sectors such as healthcare, and the most recent legislation have led to the emergence of explainability and interpretability areas.

A variety of explainability methods are presented in this section.

3.2.1 Intrinsically interpretable models

Implementing interpretable models with only a subset of relevant features is the most straightforward way to achieve interpretability. Over the years, many authors have used simple methodologies to generate human-understandable results.

Thresholding-based methods are the simplest and most interpretable ones. Usman et al. [70] have proposed a simple seizure prediction algorithm applying a threshold on the extracted univariate features. Based on the obtained results, it was possible to identify the start of the pre-ictal state in the patient's EEG recordings. Aarabi et al. [65] have also applied a simple thresholding procedure to the time profile of the extracted features. This strategy aims to determine significant modifications in the values of the selected features in comparison with a baseline, described as a reference period remote in time from any seizure.

Pinto et al. [85] proposed a simple Evolutionary Algorithm to search for discriminative features considering the best trade-off between seizure prediction performance and patient discomfort. The methodology provides patient-specific interpretable insights that could lead to an improved understanding of seizure generation processes and the underlying decisions made by the algorithm. Furthermore, a logistic regression classifier, an intrinsically interpretable model, was used in the classification step. Despite the apparent complexity of the Evolutionary Algorithm, the final result is an interpretable classifier applied to a small subset of features.

In logistic regression, the interpretation of the coefficients gives insights into which features are the most discriminative. Hence, it is necessary to transform the coefficients (natural logarithms of the odds ratio) into the odds ratio. After the transformation, the odds ratio can be analyzed to find out which features are the most important: while an odds ratio with a value around 1 is the less influential, a higher or lower one has more impact on prediction.

3.2.2 Non-interpretable models

3.2.2.1 Explainability in Epilepsy studies

This section presents not only explainability methods in seizure prediction studies but also in seizure detection and seizure classification. This search was carried out using Google Scholar and the following search strings: "explainability in seizure prediction" and "interpretability in seizure prediction". Most of the encountered research studies on this topic are related to deep learning models, not implemented in this thesis. Table 3.5 summarizes the research results.

Deep learning models have lately emerged as a state-of-the-art tool in epilepsy studies. Such models can compute complex nonlinear representations of EEG signals, achieving higher performances when compared to algorithms that rely on handcrafted features. However, due to the lack of the medical/biological meaning of the extracted features and its "black-box" nature, DL models may be poorly trusted

by clinicians. Consequently, it has led to a growing interest in the explainability of deep neural networks in epilepsy studies.

Observing Table 3.5, it is possible to note the prominence of the attention mechanisms in the explainability of DL models. Attention is a powerful concept introduced in the deep learning field that attempts to reproduce cognitive attention, which permits humans to focus on a few relevant aspects when large amounts of information are being processed. As a network architecture component, it aims to enhance the most relevant part of the input data. Consequently, besides improving the model's performance, the attention mechanisms constitute a valuable tool to produce trustworthy, and human-understandable explanations for the model's predictions [89–91].

Priyasad et al. [92], Baghdabi et al. [93], and Zhang et al. [94], proposed interpretable deep learning techniques to classify and detect epileptic seizures using raw EEG signals. In these approaches, attention-based weighting mechanisms, operating over temporal and channel-wise data, were introduced to ensure that only salient information is passed to the classifier. Furthermore, the attention mechanism increases the interpretability of the models by enabling the exploration of the importance of each EEG channel based on the learned attention weights. Consequently, it is possible to capture the relevance of each brain region in the diagnosis of epileptic seizures. On the other hand, Hsieh et al. [95] introduced a novel modular architecture with two attention modules: variable and temporal attention modules. On the one hand, the variable attention module attributes weights to variables according to their significance in classification. On the other hand, the temporal attention module specifies the time intervals during which the variables identified by the variable attention module impact the classifier output. Mansour et al. [96], presented an attention-based deep learning algorithm for seizure detection. The weights of the attention module were used to explain the relevance of each feature on the output.

SHapley Additive exPlanations (SHAP) is a visualization tool used to interpret the individual predictions of the Machine Learning models. This method aims to explain the prediction of an instance by computing the Shapley values (see section 2.4.3.2) for each feature, representing its impact on the classification [10]. As shown in Table 3.5, SHAP is also used in some epilepsy studies. Namely, Gabeff et al. [97] used SHAP values to visualize relevant features to the ictal class by comparing the output difference between a baseline EEG signal and a given input and propagating this difference back to the input signal. In his turn, Dissanayake et al. [98] applied the SHAP explainable method to evaluate the contribution of each EEG channel.

In order to achieve high accuracy and interpretability, Wang et al. [99] proposed

an epilepsy detection framework employing rule-based classifiers. This study used an ensemble learning approach to extract comprehensible rules from the SVM model, providing understandable explanations related to the model's predictions.

Guidice et al. [100] presented a deep learning pipeline for interictal EEG discrimination of epileptic seizures vs. psychogenic non-epileptic seizures patients. To analyze the behavior of the proposed algorithm, features maps from the intermediate levels were extracted and compared with the input. Permutation Entropy (PE) was used here to inspect the separability of the extracted feature maps in the intermediate transformations and to prove its discriminative power growth throughout the network's depth. The PE is a tool for analyzing complex and chaotic systems to interpret time series behavior in the classification with deep learning models.

Naze et al. [101] created an interpretable pipeline for seizure classification by applying the permutation feature importance method (see Section 2.4.3.1). The importance of each spectral band was assessed by permuting the feature value and retraining the model. Then comparing the original and retrained models, a drop in performance indicates that features are relevant.

Dissanayake et al. [102] introduced a novel Graph Neural Network-based deep learning framework for subject-independent epileptic seizure prediction. This study used qualitative evaluations and data visualization to understand the hidden patterns that lead to the graphs generated by the deep learning models. Additionally, Tang et al. [103] proposed a graph-based modeling approach for seizure detection and classification. Occlusion-based interpretability analyses were also implemented. Occlusion maps represent the relative change in the model output when a given EEG channel is dropped. It can be crucial to determining the course of seizure treatment since it evaluates the model's capability to localize the seizure. When comparing the original and retrained models, a drop in performance indicates that features are relevant.

Thomas et al. [55] presented a deep learning-based algorithm for seizure detection. This study also aims to demonstrate the capability of the algorithm to encode signal features known from medical research to be relevant for seizure detection, such as power band features. For this purpose, a canonical correlation analysis (CCA) was computed to evaluate the correlation between the power band features of the EEG signal and the representation obtained by the deep model.

Uyttenhove et al. [104] suggested a deep learning solution for automatically diagnosing epilepsy in routine EEG data. In this research Gradient - the weighted Class Activation Mapping (Grad-CAM) technique is applied to explain the model's prediction. It is frequently used to introduce transparency, especially in CNN mod-

els, since they rely on the convolutional layers' characteristic of maintaining the spatial information present in the input data. The gradient flowing from the target class into the last convolutional layer is then used to weigh the relevance of each neuron. The result is a heat map highlighting the input data segments that positively influenced the classification.

Hossain et al. [105] proposed a deep CNN model for seizure detection, able to extract spectral and temporal features from EEG data. This study uses interpretation techniques to evaluate which features from which layers are used by the CNN model. Thus, correlation maps are used to analyze how the algorithm learns from spectral amplitude features. The correlation maps were introduced by Schirrmeister et al. [106] to understand how CNNs learn to solve different tasks. This study aims to understand how Convolutional Networks of distinct architectures can be designed and trained for end-to-end learning of EEG recorded in human patients, and how appropriate visualization methods can improve their interpretability.

Cook et al. [24] developed the first-in-man study to evaluate the clinical feasibility of a long-term implanted seizure advisory system created to predict seizure likelihood. The sixteen best-performing features during the training were used as input for classification. The employed algorithm was based on characteristics of both a decision tree and a k-nearest neighbor (kNN) classifiers, which are intrinsically interpretable. However, a given feature vector could be classified through a collection of decision surfaces, which resulted in a division of the feature space into 2^{10} partitions. Therefore, the classifier cannot be considered intrinsically interpretable due to the high dimensionality of the feature space.

Decision trees are like a rule system, starting in the root node and following the path for a record to a leaf node where it is possible to see the prediction. In tree-based models, the data is split according to specific cutoff values in the features. The interpretation of the tree structure is undoubtedly simple since it has a natural visualization, with nodes and edges, easier to understand than points on a multi-dimensional hyperplane, creates good explanations, and is optimal for capturing interactions between features in the data. It is more challenging to interpret when the tree has a higher depth, being humanly incomprehensible from a given depth [10].

On the other hand, the k-nearest neighbor method uses the nearest neighbors of a data instance for prediction, assigning to the data point the most common class. Hence, this model is an instance-based learning model. It is only necessary to retrieve the k neighbors used for the prediction to provide good explanations. An instance, however, may not be interpreted if it contains hundreds or thousands of

features [10].

Moghaddam et al. [107] proposed a robust algorithm to predict seizures based on the assumption that the coherence across the brain regions changes from interictal to preictal. Studying these changes can indicate a pattern to distinguish between these two states. Therefore, in contrast to neural networks, this work provided high performance without compromising interpretability.

3.2.2.2 Explainability in other EEG studies

In addition to the above-presented works, some research was done in order to identify the explainability methods used in other studies related to EEG decoding or classification.

A brief overview of the encountered studies is presented in this section and is summarized in Table 3.5.

To explain a CNN implemented for automated sleep stage classification, Ellis et al. [108] introduced a novel local spectral explainability approach. This method was applied to demonstrate how the importance of frequency bands changes over time and to provide spectral insight into a classifier trained on raw EEG data. The data samples were converted to the frequency domain, perturbed, and back to the time domain. The percent change in classification probability from the original to the modified data was calculated for each sample to evaluate the importance of the frequency bands.

Briden et al. [109] presented a multimodal fusion architecture to classify neurological events and determine regions of the brain that influence the model's prediction. The WaveFusion SE employs Lightweight CNN (LWCNN) sub-models, trained independently for extracting localized time-frequency features, and Squeeze and Excitation Network (SEN), an attention module used to classify and identify the influential regions of the brain in the final classification.

Andreotti et al. [110] proposed a simple CNN framework for automated sleep scoring using Continuous Wavelet transformed EEG, EOG and EMG recordings as input. Further, Guided Gradient-weighted Class Activation Maps (Guided Grad-CAM) were applied to provide insights into the networks' classification mechanisms. This approach is a valuable tool for determining the relevant features in the model's prediction since it can generate fine-grained activation maps in the time-frequency domain for each signal.

Attention mechanisms are also frequently applied in other EEG studies. Jia et al. [111] presented the SalientSleepNet, a multimodal salient wave detection deep network for sleep scoring. This model integrates a multi-scale extraction module

to capture transition rules among sleep stages and a multimodal attention module to identify the relevant features for certain sleep stages. On the other hand, Phan et al. [112] introduced the SleepTransform network, an automatic sleep staging model. Furthermore, this approach addresses the interpretability of sleep staging by evaluating the scores of the self-attention module.

Vilamala et al. [113] proposed a framework capable of accurately classifying sleep stages and providing a visual interpretation of the model's prediction. Sensitivity maps are calculated to identify the most influential features in the network, providing highly interpretable images of the model behavior. Analyzing the generated maps can provide interpretable patterns that enhance communication and interaction with the domain experts.

Al-Hussaini et al. [114] presented an interpretable sleep staging model (SLEEPER), which combines deep learning architectures with expert-defined rules, specifically sleep scoring rules, via a prototype learning framework. This approach produces simple interpretable models such as logistic regression and decision trees.

3.3 Summary

The general seizure prediction framework comprises several steps, including EEG collection, signal pre-processing, feature extraction, feature selection and classification, followed by a regularization step and performance evaluation. Despite the existence of this general pipeline, there is a great variety of approaches due to the application of different methods and parameters. Regardless of reaching encouraging results, current seizure prediction approaches present numerous issues which should not be neglected.

The databases represent one of the most prominent limitations since they are mostly collected from patients during pre-surgical monitoring, which does not reflect real seizure activity. Long-term EEG recordings, comprising several months or years, acquired in an everyday routine, represent a step forward in the clinical viability of the designed methodologies.

On the other hand, the lack of proper evaluation using relevant metrics, such as sensitivity and FPR/h, as well as the absence of statistical validation, is prevalent in numerous studies. These problems, along with the results presented only for the optimal SOP and not for a range of values, cause bias and make comparisons among studies difficult.

The lack of interpretability and explainability of the seizure prediction ap-

Table 3.5: Studies associated with explainability in Epilepsy and EEG decoding and classification.

Author, year	Problem in study	Data type	Classifier	Explainability Methods
Wang et al., 2016 [99]	Detection of Epilepsy	EEG	SVM, RF, C4.5, SVM+RF, SVM+C4.5	Rule-based explanations
Schirrmeister et al., 2017 [106]	EEG Decoding and Visualization	EEG	Deep, Shallow, Hybrid and Residual ConvNet	Network correlation maps (visualization)
Vilamala et al., 2017 [113]	Automated Sleep Stage Scoring	EEG	CNN (VGGNet), Transfer learning	Sensitivity maps
Andreotti et al., 2018 [110]	Automated Sleep Stage Scoring	PSG	CNN	Guided Gradient-weighted Class Activation Maps (Guided Grad-CAM)
Al-Hussaini et al., 2019 [114]	Automated Sleep Stage Scoring	PSG	Deep prototype learning method	Interpretable models
Hossain et al., 2019 [105]	Seizure Detection	EEG	Deep CNN model	Network correlation maps (visualization)
Uyttenhove et al., 2020 [104]	Detection of Epilepsy	EEG	Tiny Visual Geometry Group (t-VGG) CNN	Gradient-weighted Class Activation Mapping (Grad- CAM)
Thomas et al., 2020 [55]	Seizure Detection	EEG	DNN (Bottleneck Network Architecture)	Analysis of latent features
Zhang et al., 2020 [94]	Seizure Detection	EEG	Adversarial learning framework	Attention mechanisms
Mansour et al.,2020 [96]	Seizure Detection	EEG	CNN, BiLstm, FCNN	Attention mechanisms
Hsieh et al., 2021 [95]	Seizure Detection	EEG	Explainable Convolutional Attention network	Attention mechanisms
Baghdadi et al., 2021 [93]	Seizure Detection and Classification	EEG	Attention-based Deep LSTM	Attention mechanisms
Priyasad et al., 2021 [92]	Seizure Classification	EEG	Deep learning architecture with attention-driven data fusion	Attention mechanisms
Dissanayake et al., 2021 [98]	Seizure Prediction	EEG	CNN, Siamese network	SHAP
Naze et al., 2021 [101]	Seizure Classification	EEG	SVM (linear and RBF), RF and Decision tree	Feature importance
Gabeff et al., 2021 [97]	Seizure Detection	EEG	CNN	Gradient ascendent, SHAP
Jia et al., 2021 [111]	Automated Sleep Stage Scoring	PSG	U^2 structures, Multi-scale extraction and multimodal attention modules, Segment-wise classification	Multimodal attention module
Ellis et al., 2021 [108]	Explain a CNN trained for sleep stage classification	EEG	CNN	Novel Local Spectral Explainability Approach
Briden et al., 2021 [109]	Classifying subjects' anxiety levels	EEG	Squeeze-and-Excitation network	Attention scores
Phan et al., 2022 [112]	Automated Sleep Stage Scoring	EEG	Seq2Seq model	Attention scores
Giudice et al., 2022 [100]	Discrimination of Subjects with Epileptic Seizures vs. Psychogenic Non-Epileptic Seizures	EEG	CNN	Permutation Entropy
Dissanayake et al., 2021 [102]	Seizure Prediction	EEG	GCL (Geometric Deep Learning)	Graph visualization
Tang et al., 2022 [103]	Seizure Detection and Classification	EEG	GNN (Graph Neural Network)	Occlusion maps
Moghaddam et al., 2022 [107]	Seizure Prediction	EEG	SVM	Spatial coherence

proaches also constitutes an obstacle to the clinical applicability of the proposed prediction methodologies and intervention devices. Over the past years, several studies have been surging to solve the absence of explainability in machine learning approaches.

Explainability methodologies range from the simplest intrinsically interpretable models such as thresholding procedures, readily understood by humans, to more complicated models such as the SVM classifier, which already requires some methodology to support its interpretation. The most intricate ones, such as deep learning models, require specific methods due to their complexity and "black box" nature.

Present strategies are mainly directed to deep learning techniques, which try to provide brain region significance, patterns of brain connectivity, and correlations with band-waves activity. However, more accessible and understandable explanations for clinicians are necessary.

On the other hand, despite recent advances in the explainability field, the explanations lack a formal evaluation. They must be evaluated according to the trust and knowledge they transmit to the domain experts.

Significant developments in the explainability field were made for general problems. However, more specific studies and explorations are required to develop proper and suitable explanations for seizure prediction problems, considering their complex nature.

4

Methodology

This chapter describes the followed steps concerning the development of explanations for epileptic seizure prediction algorithms. First, an overview of the entire pipeline is presented in Section 4.1. Then, the developed seizure prediction model is described in Section 4.2. Finally, the steps undertaken to produce and evaluate explanations regarding the model's decision are reported in Section 4.3.

4.1 Pipeline Overview

The present work aims to develop explanations of EEG-based algorithms for epileptic seizure prediction to increase trust in the models' decisions and decrease the skepticism of clinicians regarding the application of such models in healthcare.

For this, a patient-tailored algorithm for epileptic seizure prediction was designed based on the most common framework presented in the state-of-the-art. After evaluating the constructed model, several explanations were created to explain the models' decisions. Afterwards, the produced explanations were tested and validated. This process was grounded on five lessons from a previous work developed by the local research team.

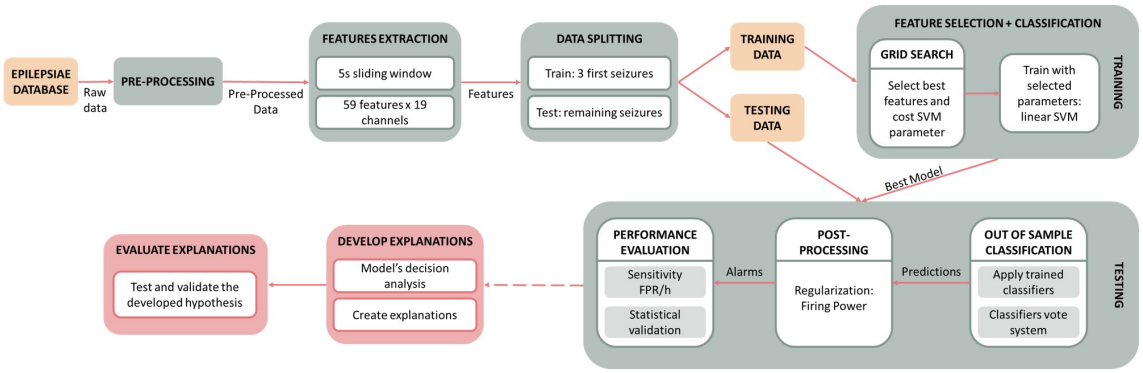

Figure 4.1: General overview of the proposed pipeline.

4.2 Seizure Prediction

4.2.1 Data

For the present study, 40 Drug-Resistant Epilepsy (DRE) patients (17 females and 23 males, with a mean age of 41.4 ± 15.7 years) were selected from the European Database on Epilepsy (EPILEPSIAE). The selected EEG data was collected by the University Medical Center of Freiburg, in Germany, from patients containing seizures localized in the temporal lobe. The data consists of EEG scalp recordings acquired with a sampling rate of 256Hz during pre-surgical monitoring. It covers 19 EEG electrodes placed according to the International 10-20 System with the following channels: FP1, FP2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, Fz, Cz, and Pz. Table 4.1 presents information about each patient (gender and age) and their seizures (number of seizures, seizure classification, seizure activity pattern, state of vigilance at seizure onset and recording time).

The selection of the 40 patients was based on the number of independent seizures. Only patients with at least four lead seizures, separated by at least 4 hours, were selected to avoid analyzing clustered seizures. As a result, 224 of 375 seizures were considered suitable for analysis.

Table 4.1: Information for the 40 studied patients.

Patient ID	Age	Sex	Number of seizures (train/test)	Seizure classification	Seizure activity pattern	Vigilance at seizure onset	Recording duration (h)
402	55	f	3	FOIA, FBTC, FOIA	t, t, t	A, A, A	103.81
			2	FBTC, FOIA	t, t	A, A	29.66
8902	67	f	3	UC, FOIA, FOIA	a, b, a	A, A, A	133.91
			2	FOIA, FOIA	m, a	A, A	22.5
11002	41	m	3	UC, FOIA, FOIA	?, s, a	A, R, A	97.16
			1	FOIA	t	A	11.7
16202	46	f	3	UC, FBTC, UC	r, ?, r	A, A, A	201.32
			4	FOIA, FOIA, FOIA, FOIA	r, r, ?, r	A, A, A, A	34.45
21902	47	m	3	UC, FOIA, FOIA	t, t, t	A, A, A	67.08
			1	FOIA	b	R	9.76
23902	36	m	3	FOA, FOA, FOA	t, t, t	A, A, A	70.74
			2	FOA, FOA	d, t	A, A	33.95
26102	65	m	3	FOIA, FOIA, FOIA	m, t, t	A, A, A	60.65
			1	FOIA	t	A	22.58
30802	28	m	3	FOA, FOA, FOA	t ,t ,t	R, A, 2	87.57
			5	FOA, FOA, FOA, FOA, FOA	t, t, t, t, t	A, A, R, 2, 2	61.71
32702	62	f	3	FOIA, FOIA, FOIA	t ,t ,t	A, A, A	117.38
			2	FOIA, FOIA	r, a	A, A	20.49
45402	41	f	3	FOIA, FOIA, FOA	t, t, t	A, A, A	71.98
			1	FOIA	t	A	22.31
46702	15	f	3	FOA, FOIA, FOIA	a, a, t	A, 2, A	47.46
			2	FBTC, FOIA	b, t	2, A	12.6
50802	43	m	3	FOIA, UC, UC	t, t, t	A, 2, 2	165.93
			2	FOIA, FBTC	t, t	2, A	35.6
52302	61	f	3	UC, FOA, UC	?, ?, d	A, A, 1	76.45
			1	UC	t	A	6.85
53402	39	m	3	FOA, FOA, FOA	?, ?, ?	A, 2, A	70.31
			1	FOIA	t	A	13.73
55202	17	f	3	FOIA, FOIA, FOA	t, d, t	A, A, A	47.05
			5	UC, UC, FOA, UC, FOIA	t, t, t, r, r	A, A, A, A, A	65.37

Table 4.1: Information for the 40 studied patients.

Patient ID	Age	Sex	Number of seizures (train/test)	Seizure classification	Seizure activity pattern	Vigilance at seizure onset	Recording duration (h)
56402	47	m	3	UC, UC, UC	t, ?, ?	A, A, A	184.22
			1	FBTC	a	A	20.25
58602	32	m	3	FOIA, FOIA, FOIA	r, t, t	A, R, A	96.94
			3	FOIA, FOIA, FOIA	r, r, t	A, A, 2	23.34
59102	47	m	3	FOA, FOIA, FOIA	?, t, t	A, A, A	65.83
			2	FOIA, FOA	t, t	A, A	82.22
60002	55	m	3	FOIA, FOIA, FOIA	d, c, t	1, A, A	208.11
			3	UC, FOIA, FOIA	t, d, d	R, R, 1	152.4
64702	51	m	3	FOA, FBTC, FBTC	?, m, t	A, A, A	75.91
			2	FBTC, FBTC	t, t	A, 2	31.59
75202	13	m	3	FOA, FOA, UC	t, t, t	2, 2, A	100.94
			4	FOA, FOA, FOA, FOA	t, t, ?, t	A, A, A, A	52.63
80702	22	f	3	FOIA, FOIA, UC	b, b, ?	A, A, A	49.4
			3	FOIA, FBTC, FOIA	c, c, c	A, A, A	29.55
85202	54	f	3	FOIA, FOIA, UC	m, c, m	2, A, A	53.49
			2	UC, UC	m, m	A, A	20.42
93402	67	m	3	FBTC, FOIA, FOIA	t, t, t	2, 2, 2	98.0
			2	UC, UC	t, t	2, 2	54.07
93902	50	m	3	FOA, FOIA, FBTC	t, t, d	A, A, 2	370.83
			3	FOIA, FOIA, UC	d, d, d	A, 2, A	20.29
94402	37	f	3	FOA, UC, FOIA	?, d, b	A, A, A	120.23
			4	UC, FOA, UC, FOA	t, ?, b, ?	2, A ,2, A	30.37
95202	50	f	3	FBTC, FOIA, FOIA	b, b, b	2, 2, 2	57.6
			4	FOIA, UC, FOIA, UC	m, b, b, t	2, 2, 2, 2	89.53
96002	58	m	3	FOIA, FOIA, FOIA	t, t, t	A, A, A	48.4
			4	FOIA, UC ,FOIA, FOIA	d ,a ,t ,a	A, A, A, A	82.2
98102	36	m	3	FOA, UC, UC	?, ?, ?	A, A, A	108.61
			2	UC, FBTC	?, ?	A, A	45.68
98202	39	m	3	FOIA, FOIA, FOIA	t, a, t	A, A, A	111.33
			5	FBTC, FOIA, FOIA, FOIA, UC	t, t, t, t, t	A, A, A, A, A	49.88
101702	52	m	3	FOIA, FOIA, FOIA	t, t, t	A, A, A	28.41
			2	FOIA, FOIA	r, r	2, A	23.83
102202	17	m	3	FOA, UC, FOIA	b, ?, t	2, A, 2	57.45
			4	UC, FOA, FOIA, UC	?, t, t, t	A, A, 2, A	51.41
104602	17	f	3	FOIA, FBTC, FBTC	t, a, t	A, 2, 2	87.87
			2	FBTC, UC	t, d	2, 2	15.25
109502	50	m	3	FOIA, FOIA, UC	t, t, t	A, A, A	76.8
			1	UC	t	A	41.94
110602	56	m	3	FOIA, FOIA, FOIA	t, t, t	A, A, A	89.63
			2	FOIA, FOA	t, t	A, A	25.92
112802	52	m	3	UC, FOIA, UC	t, t, t	A, A, A	71.58
			3	FOIA, FOIA, UC	t, t, t	A, A, A	111.5
113902	29	f	3	UC, FOIA, FOIA	t, d, t	A, A, 2	61.98
			3	FOIA, UC, FOIA	t, t, t	A, 2, A	22.73
114702	22	f	3	FOIA, FOIA, UC	t, t, t	A, A, A	68.39
			5	FOIA, FOIA, FOIA, FOIA, FOIA	t, d, t, d, t	A, A, A, A, A	34.04
114902	16	f	3	FOA, FOIA, FOIA	s, b, s	A, A, A	26.55
			4	FBTC, UC, FOIA, FOIA	t, r, a, t	2, A, A, A	50.66
123902	25	f	3	FBTC, FBTC, FOIA	t, t, t	2, 2, R	152.11
			2	FOIA, FOA	t, t	A, A	30.15

Gender: female (f), male (m); Seizure classification: unclassified (UC), Focal Onset Aware (FOA), Focal Onset Impaired (FOIA), Focal to Bilateral Tonic-Clonic (FBTC); Seizure activity pattern: unclear (?),rhythmic sharp waves (s), alpha waves (a), rhythmic delta waves (d),rhythmic theta waves (t), rhythmic beta waves (b), repetitive spiking (r), cessation of interictal activity (c), amplitude depression (m); Vigilance state: awake (A), REM sleep stage (R), Non-REM sleep stage I (1), Non-REM sleep stage II (2).

4.2.2 Pre-processing

The EEG data used in the present study was pre-processed using an EEG artifact removal model based on deep convolutional neural networks (DCNN). The model was proposed by Lopes et al. [115] to automatically and quickly remove artifacts from EEG signals, such as eye blinks, eye movements, muscle activity, cardiac activity, and electrode connection interferences, in a similar way to that

performed by experts.

This approach was developed using EEG segments manually pre-processed and labelled by experts. These segments were used to train the deep learning model in order to reproduce the experts' behavior during the manual data pre-processing. It was evaluated by comparing denoised portions with the target segments. Experimental results suggested that the proposed model was able to attenuate the influence of the artifacts in the EEG signals without human intervention, making it suitable to be employed in long-term real-time scenarios such as epileptic seizure prediction. Moreover, the fact that the used data were long-term EEG recordings from epileptic patients available in the EPILEPSIAE database and are included in the data used in the present work makes it a significant contribution to the current study.

4.2.3 Feature Extraction

After the pre-processing phase, the EEG signals were segmented into windows of 5s without overlap to extract relevant features from the data. The window length was selected according to state of the art in seizure prediction. A 5s window was considered adequate to characterize EEG variations since it is a reasonable window regarding the stationarity, temporal and spectral resolution.

Univariate linear features were extracted because they present a relatively lower computational power. Furthermore, all available electrodes were used since different brain areas can be involved in the seizure generation process.

As a result, 59 linear univariate features in the time and frequency domains were computed from each window on 19 EEG electrodes using a sliding window analysis. Regarding the frequency domain, the following bands were considered: delta (0.5-4Hz), theta (4-8Hz), alpha (8-13Hz), beta (13-30Hz), and four gamma sub-bands - gamma band 1 (30-47Hz), gamma band 2 (53-75Hz), gamma band 3 (75-97Hz), and gamma band 4 (103-128 Hz) [18]. The first two gamma sub-bands represent slow gamma bands, and the last two fast gamma bands. The frequency ranges of 47 to 53 Hz and 97 to 103 Hz are excluded to remove power line noises (harmonics of 50 Hz) [116].

Gamma waves are fast oscillations usually found during conscious perception. They are underestimated and not widely investigated compared to other slow brain waves once they present small amplitude and high contamination by muscle artifacts. However, gamma activity is involved in psychiatric disorders such as epilepsy [117]. An increase in gamma activity has been reported shortly before and during an epileptic seizure [118]. Notably, a review article about the occurrence of gamma

activity in epilepsy suggests that epileptic brain activity is a straightforward response to excessive growth in gamma activity [119]. Moreover, during epileptic seizures, gamma activity can be detected in human EEG whenever a muscle spasm occurs, as further evidence of this functional correlation [120].

Time-frequency domain features were also extracted by performing wavelet decomposition with the db4 mother wavelet. The extracted features are listed in the Figure 4.2. A more detailed description of each feature and its expected behavior can be consulted in Appendix A.

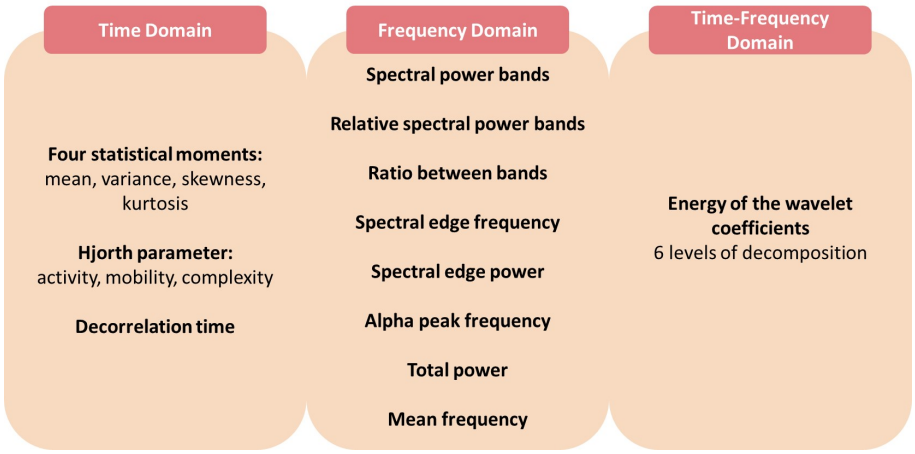

Figure 4.2: List of the linear univariate features extracted from the EEG in this work. A total of 59 features in the time and frequency domains were computed from each window on 19 EEG electrodes.

4.2.4 Data Splitting

For each patient, the feature set was split into two distinct groups: the training set, constituted by the first three seizures and used for parameters optimization and classifier training; the test set, composed of the remaining seizures and used to evaluate the classifier.

The existence of concept drifts and time dependence is assumed by using earlier seizures to train and later ones to test the models. Furthermore, the chronological division allows a real seizure prediction scenario, where the model is trained based on the initially collected seizures and then applied online to upcoming data.

As a result, 120 seizures were used in the training phase, and 104 were applied in the testing phase.

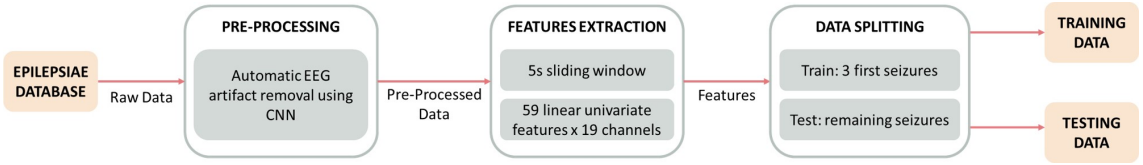

Figure 4.3: Procedure applied to prepare data for train and test.

4.2.5 Training

4.2.5.1 Class labeling

Regarding the seizure prediction problem, the samples of the feature sets were labelled into two distinct classes: preictal and interictal. The preictal class comprises the period before the seizure onset and corresponds to the total duration of the SOP and SPH.

The SPH value was set to 10 min since it is considered a suitable time interval for patients to prepare for the coming seizure. For instance, considering an alarm system in which, ideally, the patient would have time to take some medication before the seizure, such as rectal diazepam, it would take 5 to 10 minutes to work. Rectal diazepam is an anticonvulsant medication approved by the Food and Drug Administration (FDA) and used to stop a cluster of repeated seizures. Therefore, a 10-minute interval was considered appropriate for the intervention time. The samples corresponding to this period were removed from the dataset [121].

On the other hand, several values were analyzed for the SOP duration: a 10-minute minimum duration was established based on the most commonly used values in seizure prediction state of the art, and a 50-minute maximum duration was defined since patients often prefer preictal periods shorter than one-hour [122]. Therefore, several values were studied and tested for each patient: 10, 15, 20, 25, 30, 35, 40, 45, and 50 minutes.

4.2.5.2 Class balancing

As seizures are relatively rare events, there is a significant imbalance between interictal and preictal samples. A class balancing procedure was implemented during the training phase to avoid bias and specialization of the classifier over the majority class.

Therefore, systematic random undersampling was performed to obtain an equal number of samples from each class (see Figure 4.4). This method was applied in each seizure, preserving the sequential chronology of the events. During the process, the interictal set was divided into n groups, corresponding to the total

length of the preictal class, and one sample was randomly selected from each group. The methodology used allowed for handling data imbalance while maintaining the representativeness of the data.

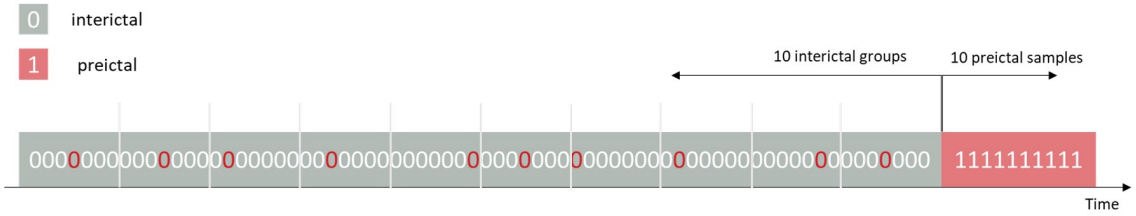

Figure 4.4: Random undersampling of interictal class respecting the sequential chronology of samples. Red colored samples correspond to interictal samples randomly chosen from each group. Only one hypothetical seizure with 10 preictal samples is illustrated.

4.2.5.3 Feature Standardization

After the class balancing stage, a standardization step was performed to normalize the range of independent features extracted from the raw data. Thus, the employed method was the z-score normalization used to standardize every value in a dataset such that the mean of all values is 0 and the standard deviation is 1.

4.2.5.4 Feature Selection

The most discriminative features were selected in this step using a filter-based method. These methods choose subsets of features based on their relationship with the target. Filter methods are simple, faster, and less computationally expensive when compared with other feature selection procedures.

The metric employed was the ANOVA (Analysis of Variance) f-test that evaluates the level of linear dependency between each feature and the target. This method returns the ranking features according to the dependence degree, allowing the selection of the k most discriminative features.

Regarding the most suitable number of features to be selected (k), a grid-search procedure was applied to tune this parameter.

4.2.5.5 Classifier

The classifier used in this work was the SVM since it has been widely used in the seizure prediction literature, presenting promising results. Furthermore, this model involves a few parameters to optimize.

The kernel selected was linear as it is simpler and computationally lighter, showing similar performances to more complex ones [66]. The parameter C (cost) controls the trade-off between smooth decision boundaries and correct training point classification [65]. A high value of C generates more intricate decision curves to fit all the points, which may lead to overfitting. The C parameter was tuned using a grid-search procedure to achieve a balanced curve.

Due to the stochasticity intrinsic to the random undersampling performed during the class balancing, an ensemble learning approach was implemented. In this procedure, 31 SVM classifiers were trained with different data samples. This number was selected in order to achieve statistical significance. Furthermore, it was considered an odd number of classifiers to avoid ties in the testing phase.

4.2.5.6 Grid-Search

A grid-search strategy was adopted to find the optimal parameters to train the SVM classifier. It included the search for the most suitable number of features (k), the appropriate value for the SVM hyperparameter (C), and the most suitable preictal period (SOP). For parameter k, it was considered four different values (10, 20, 30, and 40 features), and for C, eleven distinct values (2^{-10}, 2^{-8}, 2^{-6}, 2^{-4}, 2^{-2}, 2^{0}, 2^{2}, 2^{4}, 2^{6}, 2^{8}, 2^{10}). As a result, 44 combinations (k, C) were evaluated for each SOP value (defined in Section 4.2.5.1).

As illustrated in Figure 4.5, the Leave-One-Out Cross-Validation (LOOCV) strategy was implemented to find the optimal parameters. Therefore, considering the training set, two seizures were used to train the classifier, and the remaining one was utilized as the validation set to evaluate the classifier. For each combination (k, C), all training seizures were used precisely once to validate the model, resulting in three iterations of the LOOCV technique. This partitioning strategy ensures that training and validation sets contain samples from preictal and interictal classes.

A performance metric that transmits a trade-off between $\mathrm{SS}_{\mathrm{sample}}$ (Equation 2.1) and $\mathrm{SP}_{\mathrm{sample}}$ (Equation 2.2) was selected to evaluate the model: $\sqrt{SS_{sample} \times SP_{sample}}$. An ensemble learning method was also applied, and each iteration was executed 31 times. Therefore, for each combination (k, C), the final performance corresponded to the average metric value obtained for the 31 classifiers trained for all 3 LOOCV iterations.

Following the evaluation of all combinations (k, C), the one with the highest metric was selected as the optimal. Finally, an ensemble of 31 classifiers was trained using the best parameters and the entire training set (3 seizures). This procedure was applied to all SOP values, and the one with the best metric performance was

considered the final SOP. Nevertheless, the testing phase was applied to all assessed SOP values.

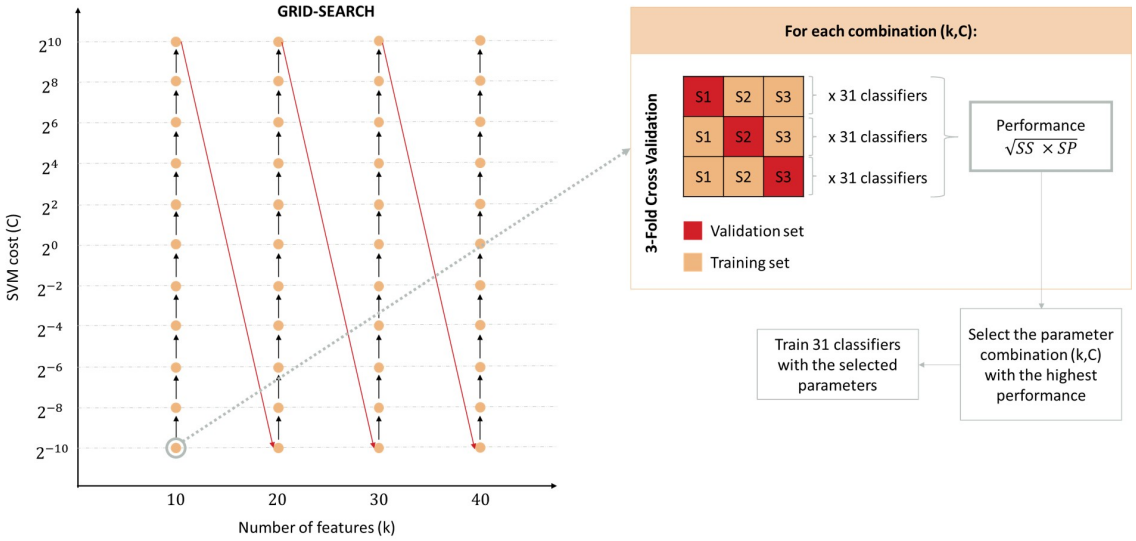

Figure 4.5: Grid-search procedure implemented to select the optimal training parameters for each preictal period.

4.2.6 Testing

After training the model, an out-of-sample classification was applied in the testing set to make predictions. As depicted in Figure 4.6, the procedure applied to the testing data was the same as the training set, excluding the class balancing. Thus, the testing set was standardized, using the z-score parameters of the training set, and the most relevant features identified in training were selected. Finally, the SVM classifier was employed to determine the output.

This procedure was executed for each of the 31 trained classifiers, resulting in 31 predictions per sample. Therefore, a voting system strategy was employed. For a given instance, the most predominant class in all predictions was assigned to the final output.

4.2.7 Post-processing

After the classification, a regularization step was performed to reduce the number of false alarms and noise and to give some connotation to the consecutive independent outcomes of the classifier, considering the output's temporal dynamics.

The selected method was the Firing Power (see Figure 4.7), computed for each epoch according to the process described in section 3.1.7. After that, an alarm was

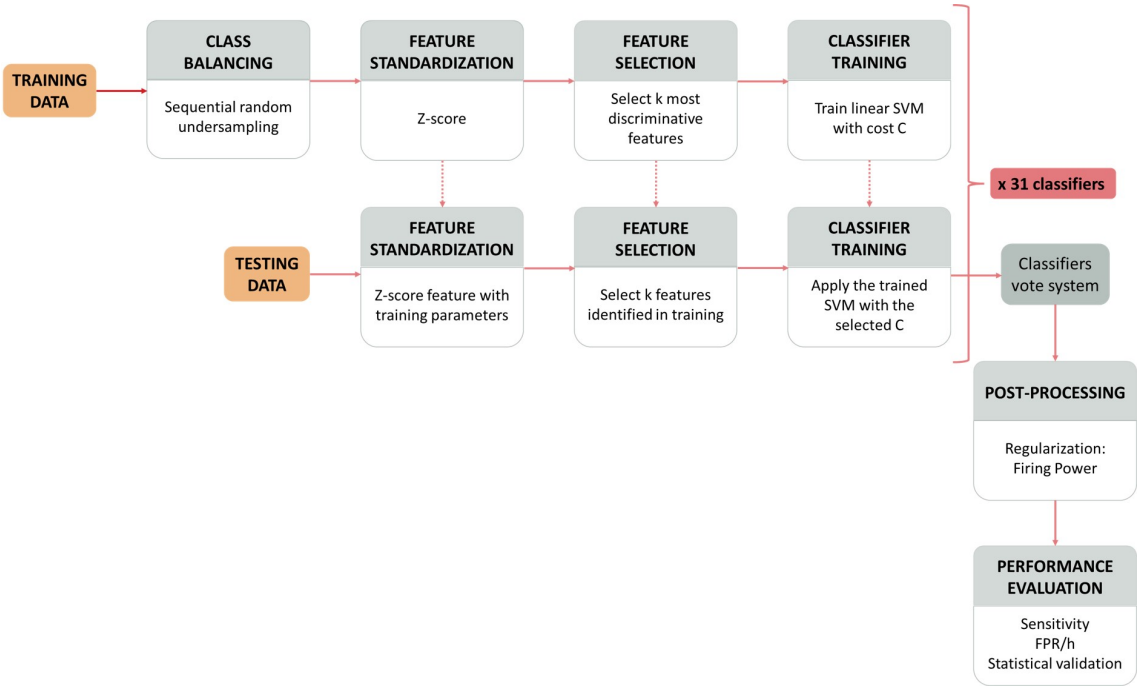

Figure 4.6: Procedure applied to train an test the seizure prediction model.

triggered when the Firing Power value exceeded a predefined threshold and was separated at least one refractory period from the last generated one. The threshold value was defined as 0.5 [63], and the refractory period corresponds to the total duration of the preictal (SOP + SPH).

The refractory period was considered in order to minimize consecutive alarms during the same seizure and reduce the patient's stress and anxiety.

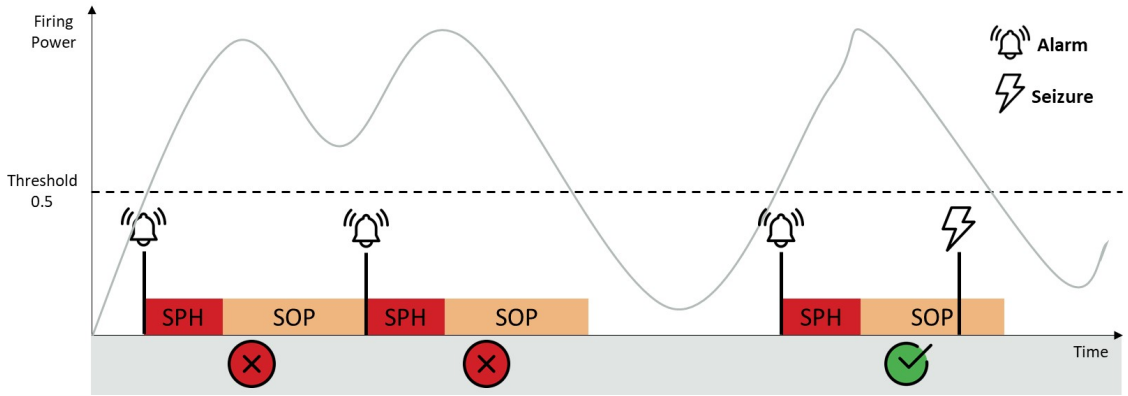

Figure 4.7: Visual representation of the firing power technique implemented. Given a certain threshold (dashed line), an alarm is only triggered when the firing power exceeds its value and is at least one refractory period separated from the last generated one. Two false alarms and one true alarm are illustrated.

4.2.8 Performance Evaluation

The performance of the seizure prediction model was evaluated using the standard metrics: sensitivity (Equation 2.3) and False Positive Rate per Hour (Equation 2.5), described in Section 2.3.2. Along with the performance assessment, a statistical validation strategy was also performed using the seizure-times surrogates method, characterized in Section 2.3.3.2.

Concerning the statistical validation, the seizure-times surrogates method was implemented to confirm if the developed algorithm performs above the chance level. This procedure randomly shifted the original onset time to another location within the interictal period. It was carried out seizure by seizure to guarantee that the artificial seizure times respect the seizure distribution over time. The surrogate times (new labels) were then used to calculate the sensitivity.

This process was executed 30 times, and the resulting average sensitivity was compared against the sensitivity calculated for the proposed methodology. The developed algorithm performs better than chance if its sensitivity is higher than the surrogate one and statistically significant. A one-sample t-test was used to evaluate this, considering a statistical significance of 0.05 under the following null hypothesis: "the sensitivity of the proposed methodology is not superior to the sensitivity of the surrogate predictor".

4.3 Explainability

The methods employed to produce explanations regarding the seizure prediction algorithm were defined based on a prior work developed by the local epilepsy research team.

4.3.1 Prior work

In the previous work, three different pipelines were developed and evaluated:

- a logistic regression;
- an ensemble of 15 Support Vector Machines;
- an ensemble of three Convolutional Neural Networks.

These methodologies were selected, considering different levels of complexity and transparency.

After evaluating each prediction methodology, patients with specific performances were selected to explain the model's decision. Thus, it was chosen patients

with: high sensitivity and very low FPR/h, high sensitivity and high FPR/h, low sensitivity and high FPR/h, and low sensitivity and low FPR/h.

Behind choosing the most suitable cases, several explanations were produced and presented to data scientists and clinicians. Concerning clinical information, some details were provided: onset times, seizure type classification, annotated EEG patterns by clinicians, and vigilance state at seizure onset.

On the feature level, it was presented explanations of methods described in section 2.4.3: beeswarm summary plots of SHAP Values, Partial Dependence Plots (PDPs), Individual Conditional Expectation (ICE), and logistic regression coefficients, when the methodology under study concerned the logistic regression. The most crucial features were also discussed regarding their expected behavior and the selected electrodes. Furthermore, calibration curves were presented to provide explanations regarding the features. A calibration curve is a visual method that evaluates how well-calibrated the classifier is.

The time-series aspect of predicting seizures was the focus of the remaining explanations. The classifier output over time was plotted along with the sleep/awake model, the interictal and preictal periods, and the raised alarms. It may allow understanding the classifier dynamics regarding concept drifts such as circadian and sleep-wake cycles. Furthermore, counterfactual explanations were provided for interesting points. They were computed by finding the slightest change in feature values that alters the prediction.

For CNN models, Local Interpretable Model-agnostic Explanations (LIME) were also used to reveal the EEG points that support the neural network window classification as preictal. This explanation was only presented to clinicians since data scientists do not have the essential background to evaluate this question.

Afterwards, the developed explanations were presented to ML experts (data scientists that work on clinical problems) and clinicians (neurologists and EEG technicians that work in an epilepsy refractory center), followed by open-ended question interviews. The answers were analyzed with the qualitative research tool, Grounded Theory (GT), which allowed for the extraction of emergent topics and ideas crucial to better understanding model explanations and their importance within the EEG seizure prediction research field. Based on these findings, five lessons were formulated to provide insights into the development of explanations.

One of the lessons highlights the importance of dividing the explanations into several levels according to a sort of granularity:

- Feature level: show and examine the obtained features, namely their signal characteristic measure, time window, and correspondent electrode;

- Model level: demonstrate how well the model distinguishes independent data samples;
- System function over time: provide a visual overview of the system output, supplying more information, namely the distribution of false alarms and firing power over time. It is also possible to provide additional information, such as the sleep/wake cycle, which might help to interpret the classifier's decisions;
- System function over specific moments: provide deeper explanations regarding strategic moments. When all recorded hours before the seizure are analyzed, it is impossible to inspect in detail all the signals. Therefore, paying more attention to false alarms, not predicted seizures, or even firing power peaks that do not lead to seizures is crucial.

It is also recommended to present the explanations according to the granularity order.

The second lesson emphasizes the importance of discussing features, namely their behavior over time and clinical details.

Another critical point is intimately related to the system function over time: time plots are the most intuitive explanations. Their inspection may reveal patterns in the model decision over time, and it is possible to formulate hypotheses as explanations to present to domain experts. These hypotheses based on physiological facts and the model's behavior should help to develop a complete problem formulation and increase robustness. It is important to note that since these explanations are mere inferences, they must be verified and tested.

The fourth finding is that it is essential to understand the differences in concepts between data scientists and clinicians. Regarding interictal and preictal concepts, the Machine Learning (ML) experts assume the existence of the preictal period as a gradual and slow transition from regular activity to a seizure. However, for clinicians, the preictal period is seen as a fast spontaneous phenomenon that might occur in a period shorter than one second, arising some seconds before the seizures.

Finally, the fifth lesson demonstrates how to explain an ML model decision to clinicians when its physiological groundings are not yet established. Therefore, this lesson states that the goal of explainability is to strategically find forms of making and testing conjectures based on the developed explanations. If they stand against these strategies, the models gain trust. If they fail, the study assumptions should be reviewed and methods redesigned, leading to a completer problem formalization. Then, new explanations are developed, and the loop continues until the models are trustworthy.

This finding is strictly associated with the legacy of Karl Popper's known falsi-

fication principle. Popper suggested that for a theory to be considered scientific, it must be able to be tested and possibly proven false if its predictions are shown to be wrong. For instance, observing a black swan can falsify the premise that "all swans are white.". According to Karl Popper, science should try to discredit a theory rather than sustain theoretical hypotheses continuously [123–125].

Furthermore, it was possible to conclude that model transparency is not one of the most crucial aspects of explaining a model decision for seizure prediction. Results also revealed that the used explainability methods are insufficient to understand brain dynamics since they produce technical and redundant information suitable for ML experts but not clinicians.

4.3.2 Adopted Methodology

The adopted methodology was grounded on the five lessons extracted from the previous work. It is believed that, for seizure prediction, the goal of explainability is not merely to explain but rather to develop hypotheses based on physiological mechanisms. Therefore, the behavior of the classifier over time was examined, and several hypotheses were formulated.

4.3.2.1 Analysis and hypotheses formulation

Firstly, based on the system function overtime lesson, detailed analyses were performed on all patients regarding the clinical information and the firing power plots over time. Several patterns were identified through this analysis, and diverse hypotheses were presented. During the plot inspection, the following aspects were evaluated:

- Classifiers with false alarms but a good regularization curve: cases in which the false alarms or firing power peaks above the alarm threshold occur near the seizure onset. It was considered a maximum period of 1h before the seizure began since this is the highest value considered during the train for the preictal period.
- Comparison with a circadian forecasting algorithm: cases in which the seizures occurred at the same daily time, and the onset time might be sufficient to predict seizures. And also patients in which seizures occurred during the night period and were correctly predicted. The night period was considered from 10 pm to 10 am. Furthermore, a seizure risk model was developed to understand the influence of the seizure onset time. This algorithm only employs the circadian seizure information and is compared with the model under study that

uses the EEG information.

- Vigilance state: cases in which the prediction model behavior was inaccurate, but it was possible to observe a sleep-wake transition in the preictal period that suggests a causality.
- Circadian-cycle influence on the classifier: cases in which it is possible to identify rhythms of false alarms occurring within similar day periods and in which the seizure may occur later, suggesting a circadian-cycle effect.
- Wrong output: cases in which the classifier output is entirely incorrect, without or with only a few raised alarms.

The selected features were also analyzed using beeswarm summary plots of SHAP Values and evaluating the relative selection frequency of each feature and electrode.

4.3.2.2 Statistical validation

After the described examination, typical model behaviors were identified, and some hypotheses were formulated. As suggested in the fifth lesson of the prior work, some strategies were employed to test and verify the proposed ideas.

One of the drawn strategies was to test the developed hypotheses employing the binomial distribution. The binomial distribution is a standard distribution that models the probability of obtaining one of two outcomes in an experiment or survey repeated multiple times.

The binomial distribution was considered since the problem under study presents two possible outcomes: success and failure, in which success is the occurrence of a given phenomenon mentioned above. It can occur in one or more patients. This method was used to evaluate whether or not the observation of an identified characteristic occurred by chance.

Regarding the problem under study, the binomial distribution was considered for testing the following null hypothesis: the phenomenon under study was observed by chance. Therefore, it was computed to obtain the probability of, by chance, observing at least x times a given pattern in n observations. The null hypothesis is rejected if the probability of x is inferior to the significance level defined. This validation aims to prove that the recognized patterns give some meaning to the classifier's behavior and are not merely results of chance.

Depending on the phenomenon in evaluation, one or two binomial distributions might be used (see Figure 4.8). Two consecutive binomial distributions are necessary if the phenomenon is evaluated for at least one seizure in each patient. On the other

hand, if the pattern is considered at the patient level, only one binomial distribution is required.

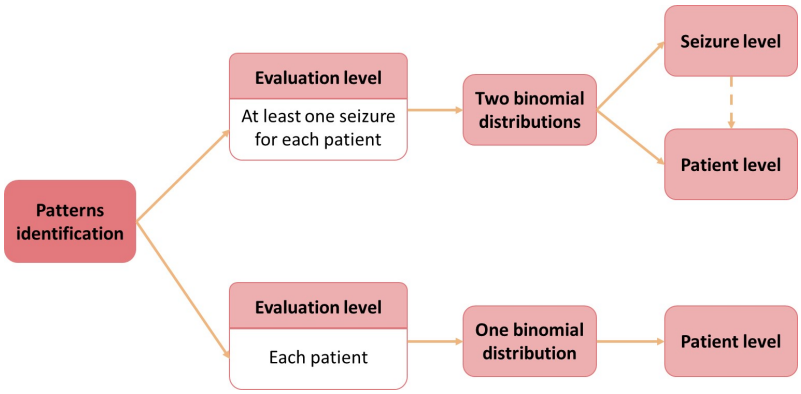

Figure 4.8: Representation of the distinction between the statistical evaluation with one and two binomial distributions.

Two binomial distributions

It was identified the patients for each at least one seizure presented one of the following cases:

- A particular seizure could not be predicted, but it is possible to trust model behavior when inspecting the firing power curve over time. It includes the cases where the true alarm was not raised due to the refractory time or where a firing power peak occurred near the seizure onset time.
- A particular seizure could not be predicted, but it was possible to observe a sleep-wake transition in the preictal period that suggests a causality.

Two consecutive binomial distributions were taken into account to evaluate these two characteristics. Two successive distributions were considered since each seizure from each patient was examined (see Figure 4.9).

Therefore, the probability of at least one seizure presenting a given factor was estimated for each patient, considering the corresponding number of testing seizures and a probability of 0.05. Then, the average value for all patients was calculated. The result shows that the likelihood that, by chance, there will be at least one seizure with a given characteristic is 0.123.

Then this probability was used to find the possibility of having, by chance, N patients (out of 40) with at least one seizure with the characteristic under study. In this situation, the complement of the cumulative binomial distribution was used, and the factor was considered statistically valid from the number of observations which have a probability of occurring by chance below the significance level considered.

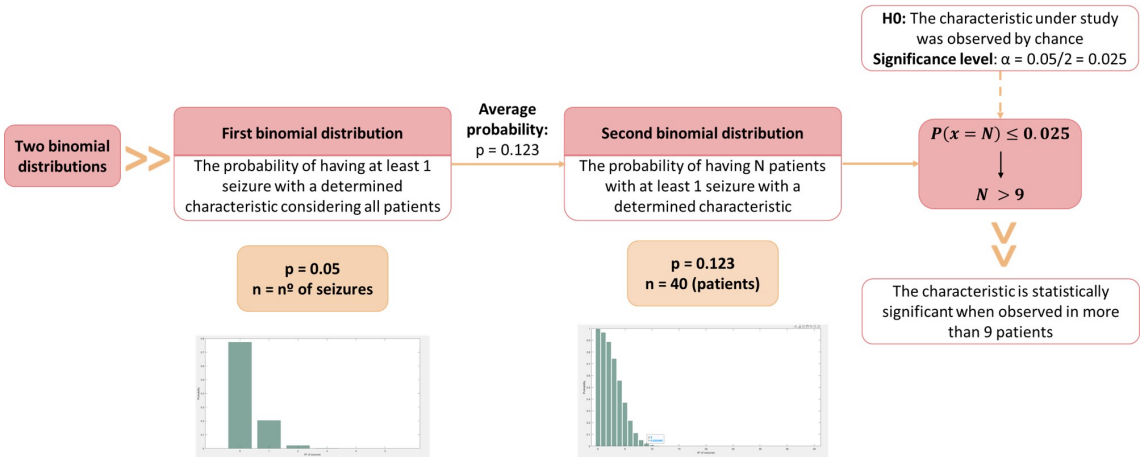

Figure 4.9: Procedure applied for statistical validation when two binomial distributions are used.

The significance level was obtained in this case based on the Bonferroni correction. Although this method is used for multiple comparisons, it was employed to turn this evaluation more rigorous as two binomial distributions were considered, and the probability for the first one was an assumption. As a result, a significance level of $\alpha = 0.05/2 = 0.025$ was used.

It was concluded that until nine patients, the observed characteristic is considered to be occurred by chance. Therefore, when the factor is observed in more than nine patients, the null hypothesis is rejected, and the phenomenon is assumed to be statistically valid.

One binomial distribution

The patients in which the circadian cycles were present were also counted and statistically validated. The same strategy was applied to the situations where the circadian forecasting algorithm performed worst and better than the seizure prediction model under study.

For these characteristics, only one binomial distribution was taken into account since it did not evaluate each seizure but the general scenario for each patient (see Figure 4.10). Therefore, the probability of having N patients (out of 40) with a determined characteristic was assessed, considering a probability of 0.05. Regarding this context, a significance level of 0.05 was used. It was figured that the characteristics are significant when the factor is observed in more than four patients.

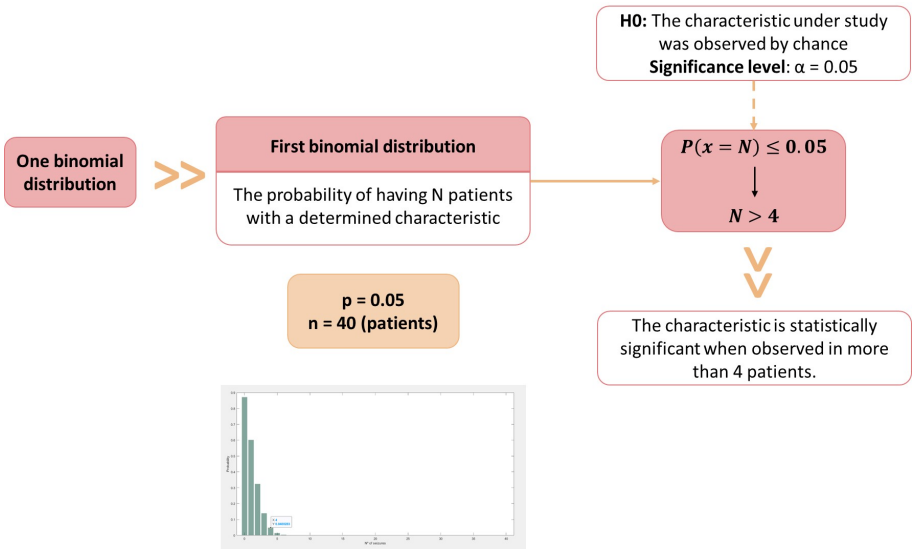

Figure 4.10: Procedure applied for statistical validation when one binomial distribution is used.

5

Results and Discussion

This chapter presents the results obtained from the proposed methodology, along with their analysis and discussion. Section 5.1 focuses on the proposed seizure prediction methodology, while Section 5.2 focuses on developing explanations and individual evaluation.

5.1 Seizure Prediction

The methodology described in the previous chapter (Section 4.2) was applied to the 40 Drug-Resistant Epilepsy (DRE) patients from the EPILEPSIAE database selected for this study.

Even though several Seizure Occurrence Period (SOP) values were considered and tested, the results presented in this section contain only one SOP duration for each patient, which was selected according to its best metric performance ($\sqrt{SS_{sample} \times SP_{sample}}$).

5.1.1 Training phase

During the training phase, a grid search procedure was implemented, and a patient-tailored model for epileptic seizure prediction was trained, considering the first three seizures of each patient.

Table 5.1 summarizes, for each patient, the selected training parameters during the grid-search procedure (SVM cost and the number of features) together with the validation results (sample sensitivity and sample specificity). Furthermore, average values of sample sensitivity and specificity considering all patients are also presented.

The results reveal a tendency to select ten features and a hyperparameter cost of 10. Furthermore, it is visible that the SOP values chosen vary considerably between patients, ranging from the minimum (10 minutes) to the maximum duration considered (50 minutes). Although, a tendency for values such as 10, 15, and 50 minutes is visible.

Table 5.1: Training parameters and performance obtained for each patient.

Patient	SOP	k	C	SS_{sample}	SP_{sample}
402	10	20	2^{-10}	0.41	0.71
8902	20	30	2^{-10}	0.87	0.84
11002	15	10	2^{2}	0.46	0.70
16202	15	40	2^{-4}	0.64	0.83
21902	10	40	2^{-10}	0.66	0.60
23902	50	40	2^{-10}	0.68	0.52
26102	50	40	2^{0}	0.32	0.62
30802	50	30	2^{-10}	0.90	0.79
32702	15	10	2^{-10}	0.76	0.69
45402	15	40	2^{-10}	0.72	0.56
46702	40	40	2^{2}	0.22	0.67
50802	15	10	2^{-4}	0.83	0.83
52302	50	10	2^{2}	0.69	0.71
53402	45	10	2^{-10}	0.48	0.67
55202	10	30	2^{0}	0.51	0.73
56402	10	10	2^{2}	0.73	0.70
58602	10	10	2^{2}	0.30	0.70
59102	15	10	2^{-10}	0.64	0.45
60002	15	10	2^{4}	0.56	0.72
64702	30	30	2^{-10}	0.42	0.68
75202	30	30	2^{-8}	0.72	0.83
80702	35	40	2^{0}	0.39	0.76
85202	15	30	2^{10}	0.45	0.67
93402	50	10	2^{8}	0.54	0.53
93902	40	30	2^{8}	0.60	0.57
94402	10	40	2^{10}	0.42	0.66
95202	10	10	2^{-10}	0.74	0.66
96002	40	10	2^{0}	0.82	0.64
98102	35	10	2^{10}	0.49	0.53
98202	10	10	2^{8}	0.29	0.70
101702	10	10	2^{4}	0.51	0.59
102202	50	10	2^{10}	0.34	0.66
104602	30	40	2^{6}	0.38	0.69
109502	10	10	2^{-8}	0.55	0.55
110602	50	10	2^{-2}	0.56	0.60
112802	10	20	2^{-10}	0.56	0.57
113902	45	30	2^{-8}	0.41	0.55
114702	35	20	2^{0}	0.29	0.70
114902	20	10	2^{4}	0.50	0.62
123902	10	40	2^{-6}	0.79	0.85
	-	-	-	0.55 $\pm$0.18	0.67 $\pm$0.10

k: number of features; C: SVM cost; SS_{sample}: sample sensitivity; SP_{sample}: sample specificity.

The average sample sensitivity and specificity values correspond to 0.55 ± 0.18 and 0.67 ± 0.10 respectively. Despite the slight discrepancy, specificity was relatively higher than sensitivity, implying a better classification of the interictal samples.

It would be expected that the classifiers would be able to better distinguish between interictal and preictal samples. However, it will only be possible to infer their real predictive power in the testing phase with unseen data. The relatively low training performance highlights the complexity of the seizure prediction problem.

5.1.2 Testing phase

During the testing phase, the developed patient-specific models were evaluated, considering the remaining seizures of each patient. Therefore, sensitivity and False Positive Rate per Hour (FPR/h) were assessed according to Equations 2.3 and 2.5, respectively. Along with the performance assessment, a statistical validation strategy was also performed using the seizure-times surrogates method.

Table 5.2 summarizes the seizure prediction results obtained for each patient.

As depicted in the table, the number of evaluated seizures varies between patients, ranging from 1 to 5. Therefore, the comparison of the sensitivity values between patients may be difficult. For instance, when only one seizure is considered, the sensitivity is limited to 0 (seizure not predicted) or 1 (seizure correctly predicted). On the other hand, several sensitivity values can be obtained when five seizures are evaluated. Therefore, a sensitivity value of 1 in a patient with one testing seizure has not the same meaning as the sensitivity value of 1 in a patient with five evaluated seizures. Indeed, observing the table, it is visible that the sensitivity value of 1 is obtained when only one seizure is considered, except for patient 8902, which has a sensitivity of 1 with two evaluated seizures. Consequently, it could be viewed as a limitation of the present study, which would only be possible to overcome with the availability of more extensive data per patient.

Regarding the classifier performance, the average sensitivity and FPR/h values obtained across the 40 patients correspond to 0.34 ± 0.35 and 1.78 ± 1.95, respectively. The relatively low average sensitivity makes it perceptible that the developed classifier has an insufficient capacity to predict seizures correctly. On the other hand, the elevated FPR/h value may lead to questioning the applicability of the developed system in real life since the high rate of false alarms per hour may bring consequences to the patient's health.

Regarding the number of predicted seizures, the patients with more true alarms were 30802 (with three predicted seizures), 8902, 55202, and 113902 (with two pre-

Table 5.2: Testing performance obtained for each patient.

Patient	Evaluated seizures	SOP	SS	FPR/h	SS Surrogate	p-value	Statistically valid
402	2	10	0.50	5.42	0.33 ±0.30	0.00	●
8902	2	20	1.00	0.16	0.12 ±0.21	0.00	●
11002	1	15	1.00	0.57	0.23 ±0.42	0.00	●
16202	4	15	0.00	0.07	0.03 ±0.08	0.96	
21902	1	10	0.00	1.43	0.17 ±0.37	0.99	
23902	2	50	0.00	2.95	0.70 ±0.36	1.00	
26102	1	50	1.00	1.49	0.43 ±0.50	0.00	●
30802	5	50	0.60	0.43	0.39 ±0.14	0.00	●
32702	2	15	0.00	0.45	0.05 ±0.15	0.96	
45402	1	15	0.00	3.40	0.33 ±0.47	1.00	
46702	2	40	0.50	1.77	0.48 ±0.38	0.41	
50802	2	15	0.00	0.43	0.08 ±0.19	0.99	
52302	1	50	0.00	2.87	0.83 ±0.37	1.00	
53402	1	45	1.00	0.87	0.47 ±0.56	0.00	●
55202	5	10	0.40	2.28	0.30 ±0.20	0.01	●
56402	1	10	1.00	5.55	0.37 ±0.48	0.00	●
58602	3	10	0.00	2.14	0.16 ±0.22	1.00	
59102	2	15	0.50	9.72	0.55 ±0.30	0.81	
60002	3	15	0.00	0.99	0.22 ±0.26	1.00	
64702	2	30	0.50	0.91	0.23 ±0.25	0.00	●
75202	4	30	0.00	0.11	0.07 ±0.11	1.00	
80702	3	35	0.33	1.31	0.29 ±0.21	0.13	
85202	2	15	0.00	0.17	0.02 ±0.09	0.84	
93402	2	50	0.50	3.88	0.80 ±0.31	1.00	
93902	3	40	0.33	0.46	0.14 ±0.22	0.00	●
94402	4	10	0.25	3.04	0.23 ±0.19	0.24	
95202	4	10	0.25	1.07	0.12 ±0.15	0.00	●
96002	4	40	0.25	2.00	0.55 ±0.26	1.00	
98102	2	35	0.50	0.25	0.12 ±0.21	0.00	●
98202	5	10	0.00	0.00	0.00 ±0.00	-	
101702	2	10	0.00	1.97	0.25 ±0.25	1.00	
102202	4	50	0.25	0.31	0.11 ±0.12	0.00	●
104602	2	30	0.50	0.50	0.20 ±0.31	0.00	●
109502	1	10	1.00	3.37	0.20 ±0.40	0.00	●
110602	2	50	0.50	1.28	0.43 ±0.31	0.13	
112802	3	10	0.33	4.39	0.29 ±0.27	0.19	
113902	3	45	0.67	2.69	0.59 ±0.22	0.03	●
114702	5	35	0.00	0.20	0.14 ±0.16	1.00	
114902	4	20	0.00	0.16	0.06 ±0.12	0.99	
123902	2	10	0.00	0.00	0.00 ±0.00	-	
Average	-	-	0.34 ±0.35	1.78 ±1.95	0.28 ±0.25	-	16

dicted seizures). Although evaluating the number of false alarms per hour, patient 8902 was the one that presented the best performance, with an FPR/h value of 0.17, which is very close to the maximum value considered adequate for a real-life application (FPR/h=0.15).

Finally, considering the statistical validation, the sensitivity of the surrogates method averaged at 0.28±0.25, where 16 patients (40%) achieved performance above the chance level.

5.1.3 Comparative analysis with other studies

Concerning the seizure prediction pipeline, the obtained results can be compared with the performances of prior studies presented in Chapter 3. With this goal, four works that employed the EPILEPSIAE database and implemented statistical validation were selected. Table 5.3 shows the performance of the best approaches of the considered studies and the developed methodology in the present work.

Table 5.3: Seizure prediction performance for studies under comparison.

Study	Number of patients	SS	FPR/h	Validated patients
Alvarado-Rojas et al. [60] (2014)	53	0.47	0.94	13.21%
Rasekhi et al. [62] (2015)	10	0.61	0.11	80%
Bandarabadi et al. [63] (2015)	24	0.76	0.1	100%
Direito et al. [66] (2017)	216	0.38	0.2	11.11%
Developed methodology	40	0.34	1.78	40%

By observing the performances of the selected studies, it is notable that superior sensitivity values were achieved compared to the developed methodology. Although, the value obtained by Direito et al. [66] was slightly more significant than in the present work. Regarding the rate of false alarms per hour, it is visible that the proposed methodology presents the highest value for the FPR/h metric, followed by the Alvarado-Rojas et al. [60] study, which showed an FPR/h value of 0.94. Concerning both metrics, Alvarado-Rojas et al. [60] is the work that presented results more similar to the developed methodology.

The random predictor was used in all four selected studies to perform the statistical validation. Bandarabadi et al. [63] presented a statistical validation of 100%, i.e. all the evaluated patients performed better than chance, and Rasekhi et al. [62] performed above chance for 80% of the patients. However, Alvarado-Rojas et al. [60] and Direito et al. [66] only attained performance above the chance level for 13.21% and 11.11% of the patients, respectively, a relatively lower percentage compared to the developed pipeline (40%).

Similarly to the proposed methodology, in the selected studies, a range value of SOP was used, and the best duration was determined for each patient. While in the present study, the selection was based on the defined training metric, in the Bandarabadi et al. [63], and Rasekhi et al. [62], this choice was made according to the best testing performance. However, it may lead to a bias in the presented results and an impeding of real-life applications once the model parameters are chosen based on the test results that are unknown a priori. Furthermore, Direito et al. [66] considered an SPH duration of 10 seconds, which is improper for a warning system since it does not provide enough time for the patient to take preventive actions.

It is worth noting that the comparison between studies is a challenging task once there is significant heterogeneity regarding the choice of the patient set and the enormous diversity of available parameters and options incorporated throughout the methodology.

5.2 Explainability

5.2.1 Time analysis

As stated before, after analyzing the metric performances, all patients' firing power plots were examined, and distinct aspects were evaluated. The detailed analysis and all patients' plots can be consulted on GitHub[1] .

Classifiers with false alarms but a good regularization curve

The number of false alarms occurring near the seizure onset was inspected since, despite false alarms, the classifier's behavior may be considered normal regarding the preictal period assumptions and its duration.

By visualizing seizure #4 from patient 8902 (see Figure 5.1), it is possible to observe that the firing power curve presents a relatively small peak at 1 pm that was far from raising the alarm. A monotonically increasing tendency follows it until reaching a maximum peak value of 1.0 in the preictal period and then decreasing. Additionally to the true alarm, a false alarm was also raised when the firing power got a value superior to 0.5. Despite being a false alarm, this behavior may be considered normal regarding the preictal period assumption: it is assumed the existence of a gradual and slow transition from regular activity to a seizure that can be captured from an EEG background analysis. Therefore, the following hypothesis can be presented: the system raised a false alarm because it may have caught the brain's slow and gradual transition from regular activity to a pre-seizure one. Several other cases were identified with a similar pattern.

In seizure #6 from patient 93902 (see Figure 5.2), despite the firing power curve presenting a value superior to 0.5 during the preictal period, a true alarm was not raised. This situation occurred due to the refractory period during which it is impossible to raise alarms. Consequently, the seizure was not predicted since it was impossible to trigger the alarm during the preictal period due to the refractory time. Furthermore, the pattern described above in the 8902 case is also visible in this patient.

In other patients, such as patient 114902 (see Figure 5.3a), despite the seizures not being predicted, a firing power peak with values superior to the alarm threshold is visible until one hour before the seizure onset. It was assumed one hour since it is the highest value considered during the train for the preictal period. These cases

[1] `https://github.com/JoanaFBatista/On-the-clinical-acceptance-of-EEG-seizure-prediction-methodologies.git`

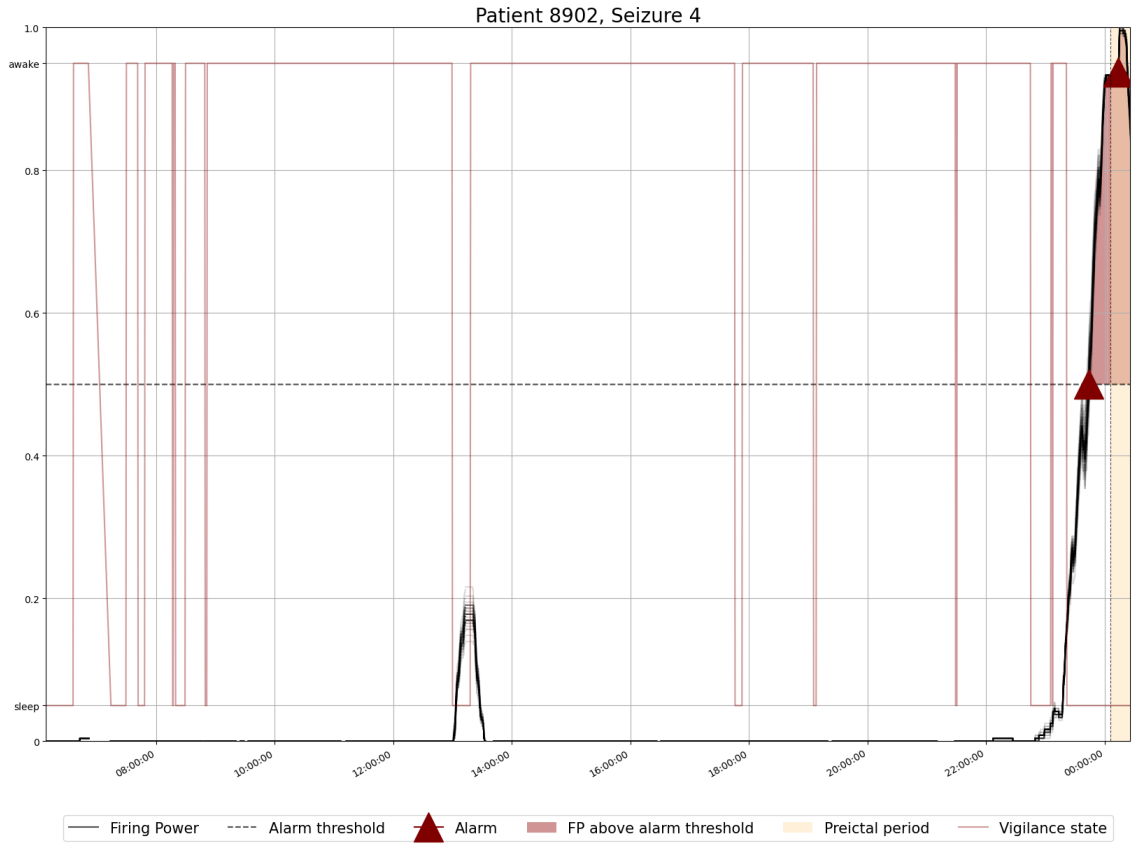

Figure 5.1: Plot of the ensemble of SVMs' decisions over time for patient 8902 seizure #4. Each SVM decision is in grey and the ensemble voting system is in black. The vigilance state is also represented.

may lead to the thought that if it was considered higher values of SOP, the seizure might be predicted. Indeed, analyzing Figure 5.3b, it is visible that if an SOP equal to or superior to 40 minutes were chosen instead of 20 minutes, the seizure would be predicted. Furthermore, evaluating the general scenario for patient 114902 (see Figure 5.3c), it is notable that if an SOP of 25 minutes were chosen instead of 20 minutes, the sensitivity would increase from 0 to 0.25. However, this growth in the number of true alarms is not linearly correlated with the increase in the SOP value. This situation is visible when a 30-minute SOP is selected, in which the sensitivity is null. It happens because the growth of SOP duration implies the augmentation of the refractory time, which means fewer alarms will be triggered per hour, and the seizure cannot be predicted.

These two last examples demonstrate how controversial the selection of the most suitable SOP value can be and the critical role it plays in classifier performance. Additionally, in both cases, despite the seizures not being predicted, it is possible to trust its behavior when inspecting the firing power curve over time.

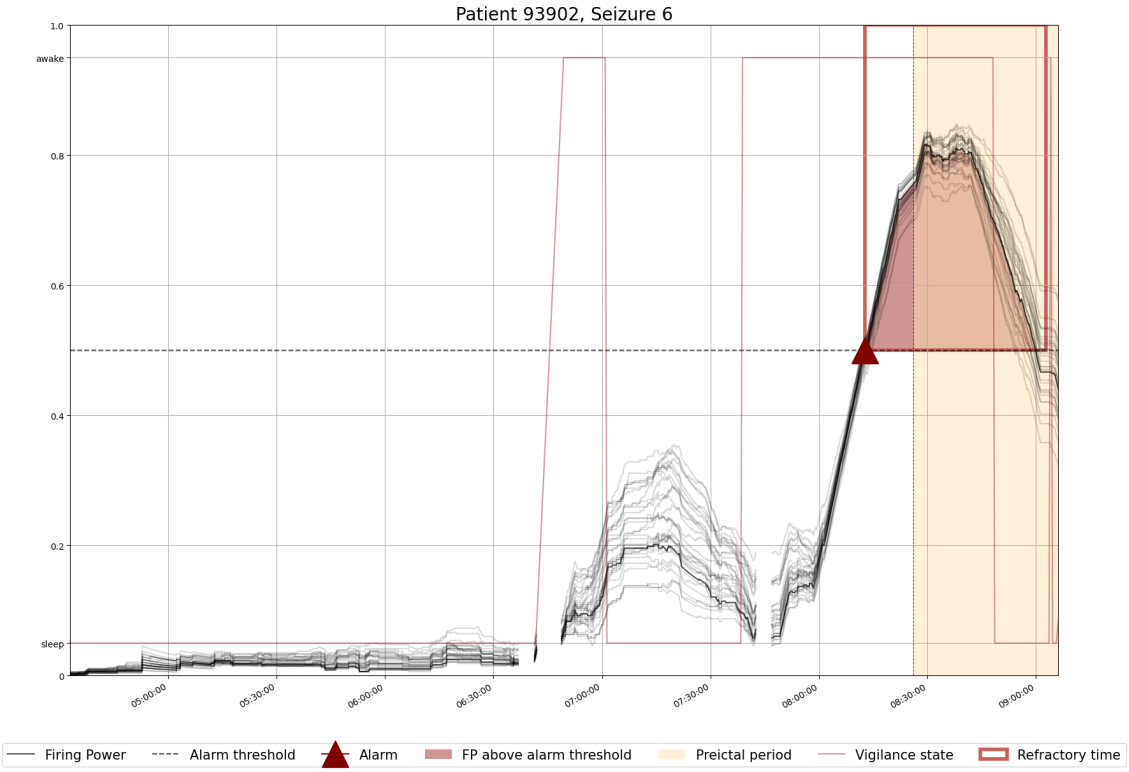

Figure 5.2: Plot of the ensemble of SVMs' decisions over time for patient 93902 seizure #6. Each SVM decision is in grey, and the ensemble voting system is in black. The rectangle drawn in the figure represents the refractory time to illustrate why no true alarm was raised during the preictal period. The vigilance state is also represented.

Comparison with a circadian forecasting algorithm

In the previous work, clinicians raised some skepticism regarding a possible bias in the classifier performance in patients in which seizures occur during the night since, in this period, the EEG is less prone to artifacts and noise. Therefore, the seizure onset time of each patient was analyzed, and those patients whose seizures happened during the night (from 10 pm to 10 am) were also evaluated. Although, only in 2 of these patients all the testing seizures were correctly predicted.

All patient seizures were also analyzed to evaluate if there was a pattern regarding the seizure onset time. It was noticed in several patients that some seizures occurred approximately at the same time. Notably, it was observed that some testing seizure onsets occur roughly at the same time as training seizures, which may lead to the question of whether the classifier has considered the training seizures onset to predict the testing seizure. However, the testing seizure was correctly predicted in only 3 of the 8 cases of this type, where the testing seizure onset coincided with a training seizure onset.

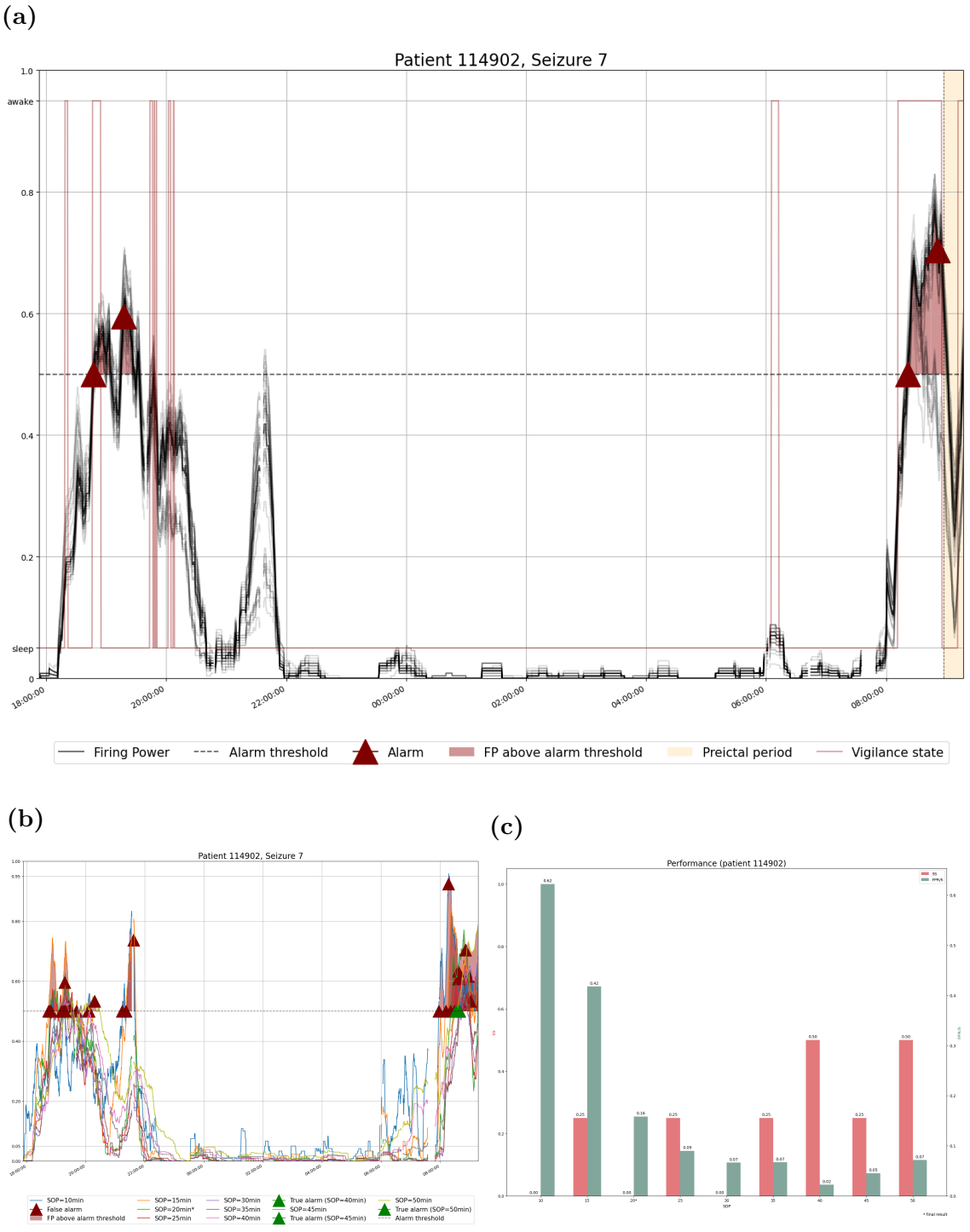

Figure 5.3: **(a)** Plot of the ensemble of SVMs' decisions over time for patient 114902 seizure #7. Each SVM decision is in grey, and the ensemble voting system is in black. The vigilance state is also represented. **(c)** Plot of the model decisions over time for patient 114902 considering all SOPs. **(c)** Testing performance (SS, FPR/h) for patient 114902 considering each SOP.

Concerning this last situation, the seizure prediction model was evaluated by comparing it with a circadian forecasting algorithm that did not use EEG data. This new model evaluates the seizure risk considering only circadian information, more precisely, the training seizures onset time. A seizure risk alarm was defined from 30 minutes before to 30 minutes after each seizure training onset time. A given testing seizure was predicted if its onset occurred during the 1-hour seizure risk interval, as illustrated in Figure 5.4.

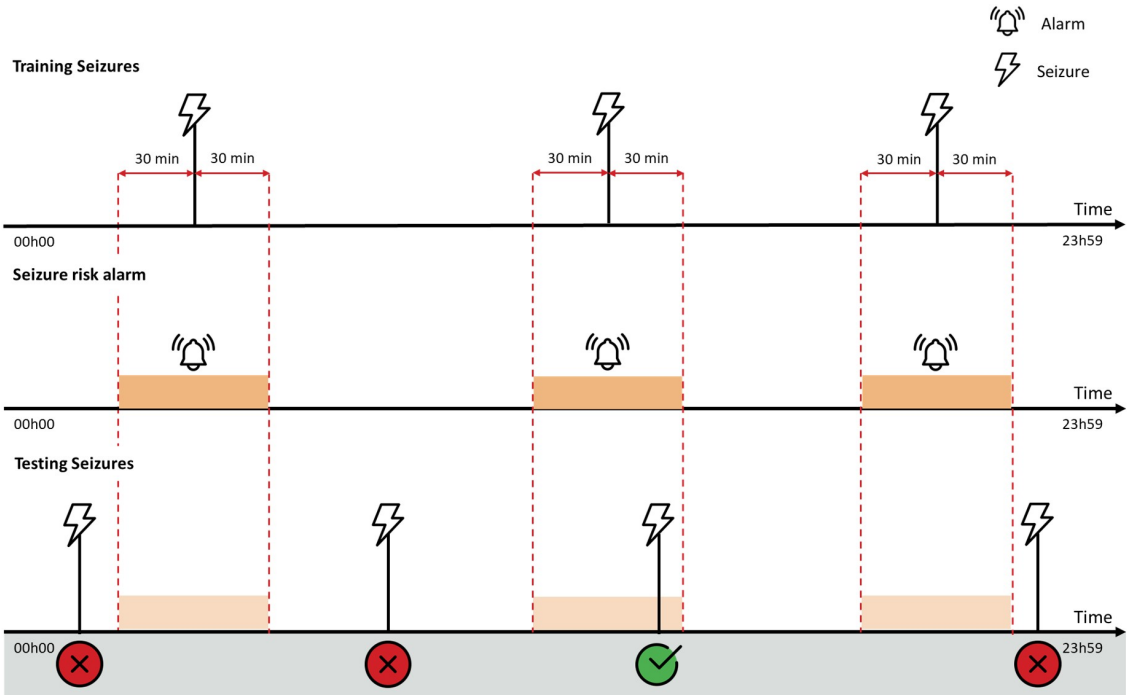

Figure 5.4: Visual representation of the circadian forecasting model implementation. One testing seizure is predicted if its onset occurs during the seizure risk interval.

The sensitivity and the time under warning were analyzed for both models, and the results are presented in Table 5.4.

The outcome showed that in four patients, the circadian forecasting model outperformed the model under study, offering a superior sensitivity value and a lower time under warning. On the other hand, despite the SVM model presenting a better sensitivity in several patients, it only performed better for three patients considering both metrics since the time under warning is generally superior for this approach. Indeed, it is notable that the circadian model presents a shorter time under warning for most cases (27 patients) than the model under study. This approach offers a significant advantage in real-life applications since it reduces the patient's stress and anxiety.

The analysis of the reported results may raise the question if there is any advan-

Table 5.4: Sensitivity and time under warning for circadian forecasting model, and respective comparison with the approach under study for each patient.

Patient	Training Seizures	Testing Seizures	Sensitivity		Time under warning		Circadian better than SVM	Circadian worse than SVM
			Circadian model	SVM model	Circadian model	SVM model		
402	3	2	0	0.50	3:00:00	9:53:15		
8902	3	2	0.50	1.00	2:47:50	1:10:20		●
11002	3	1	0	1.00	3:00:00	1:30:15		●
16202	3	4	0	0	2:31:03	0:20:55		
21902	3	1	0	0	3:00:00	2:18:30		
23902	3	2	0	0	2:59:58	18:54:15		
26102	3	1	0	1.00	2:40:55	9:21:30		
30802	3	5	0	0.60	2:18:52	17:03:55		
32702	3	2	0	0	2:09:34	1:35:35		
45402	3	1	1.00	0	3:00:00	7:41:50	●	
46702	3	2	0.50	0.50	3:00:00	5:08:15		
50802	3	2	0	0	3:00:00	4:03:40		
52302	3	1	0	0	3:00:00	2:58:20		
53402	3	1	0	1.00	3:00:00	3:08:00		
55202	3	5	0.40	0.40	3:00:00	18:12:45		
56402	3	1	0	1.00	2:56:16	9:47:50		
58602	3	3	0.33	0	3:00:00	4:34:05	●	
59102	3	2	0	0.50	3:00:00	2 days 8:49:45		
60002	3	3	0	0	2:22:06	10:25:55		
64702	3	2	0	0.50	3:00:00	10:02:50		
75202	3	4	0	0	3:00:00	1:51:20		
80702	3	3	0.33	0.33	3:00:00	10:48:45		
85202	3	2	0	0	3:00:00	0:41:20		
93402	3	2	0	0.50	3:00:00	1 day, 13:57:30		
93902	3	3	0	0.33	3:00:00	4:19:35		
94402	3	4	0	0.25	3:00:00	8:44:20		
95202	3	4	0	0.25	2:07:15	17:08:20		
96002	3	4	0	0.25	2:36:08	1 day, 13:30:40		
98102	3	2	0	0.50	3:00:00	5:44:15		
98202	3	5	0	0	3:00:00	0:00:00		
101702	3	2	0.50	0	3:00:00	6:39:35	●	
102202	3	4	0.50	0.25	3:00:00	8:09:15	●	
104602	3	2	0	0.50	3:00:00	1:42:50		●
109502	3	1	1.00	1.00	3:00:00	17:08:35		
110602	3	2	0	0.50	3:00:00	11:48:55		
112802	3	3	0	0.33	3:00:00	2 days, 4:11:25		
113902	3	3	0.33	0.67	3:00:00	13:08:00		
114702	3	5	0.40	0	3:00:00	1:15:55		
114902	3	4	0	0	3:00:00	1:45:55		
123902	3	2	0	0	2:14:27	0:00:00		
			Average SS		Average time		Relative frequency	
	-	-	0.15	0.34	02:52:06	10:50:27	0.1	0.75

tage of developing complex models using the EEG information when a simple seizure risk model that uses circadian seizure information presents a better performance for a superior number of patients.

Vigilance state

The vigilance state was also examined. In several cases, the prediction model behavior was inaccurate, but a transition in the vigilance state in the preictal period was noticed.

By visualizing seizure #6 from patient 94402 (see Figure 5.5), it is possible to observe that the firing power curve presents several peaks with values superior to the alarm threshold, generating several false alarms. Following this and right before the preictal period, the firing power curve decreases, and the seizure is not predicted. However, an awake-sleep transition is visible near the beginning of the preictal period, which suggests the presence of a causality.

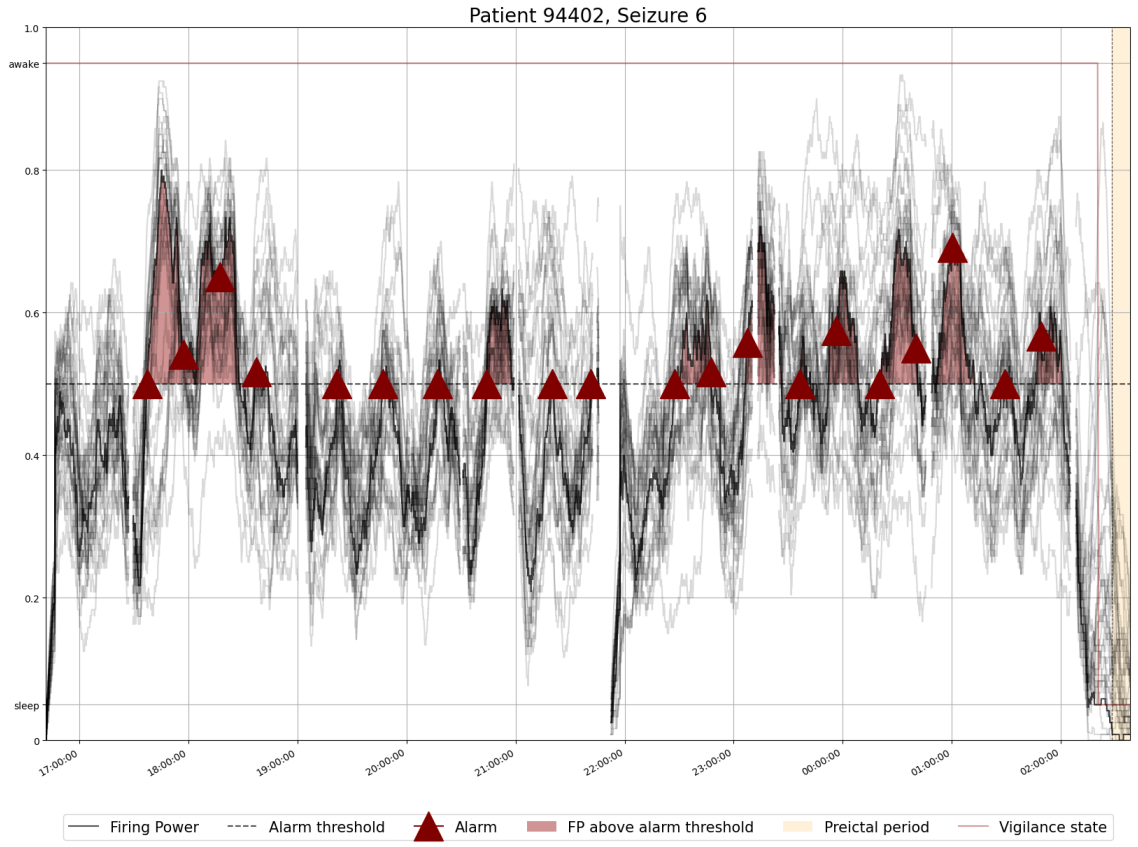

Figure 5.5: Plot of the ensemble of SVMs' decisions over time for patient 94402 seizure #6. Each SVM decision is in grey and the ensemble voting system is in black. The vigilance state is also represented.

Circadian-cycle influence on the classifier

The presence of cyclicity in the firing power curve was also evaluated. It was noticed some clusters of false alarms occurred within similar day periods, suggesting a circadian-cycle influence. It may lead to the hypothesis that these false alarm clusters may indicate the existence of periods of brain susceptibility to seizures, which may not occur. Or on the other hand, some daily actions may generate distinct EEG patterns that can be recognized as preictal activity by the classifier, causing false alarms.

Observing the firing power curve of patient 112802 (see Figure 5.6), it is possible to identify several clusters of false alarms within the same day period from 6-8 am to 10-11 pm. Considering this daily period, it is possible to speculate that such false alarm clusters may occur when the patient wakes up and falls asleep, which might cause distinct EEG patterns that can be identified as preictal activity by the classifier.

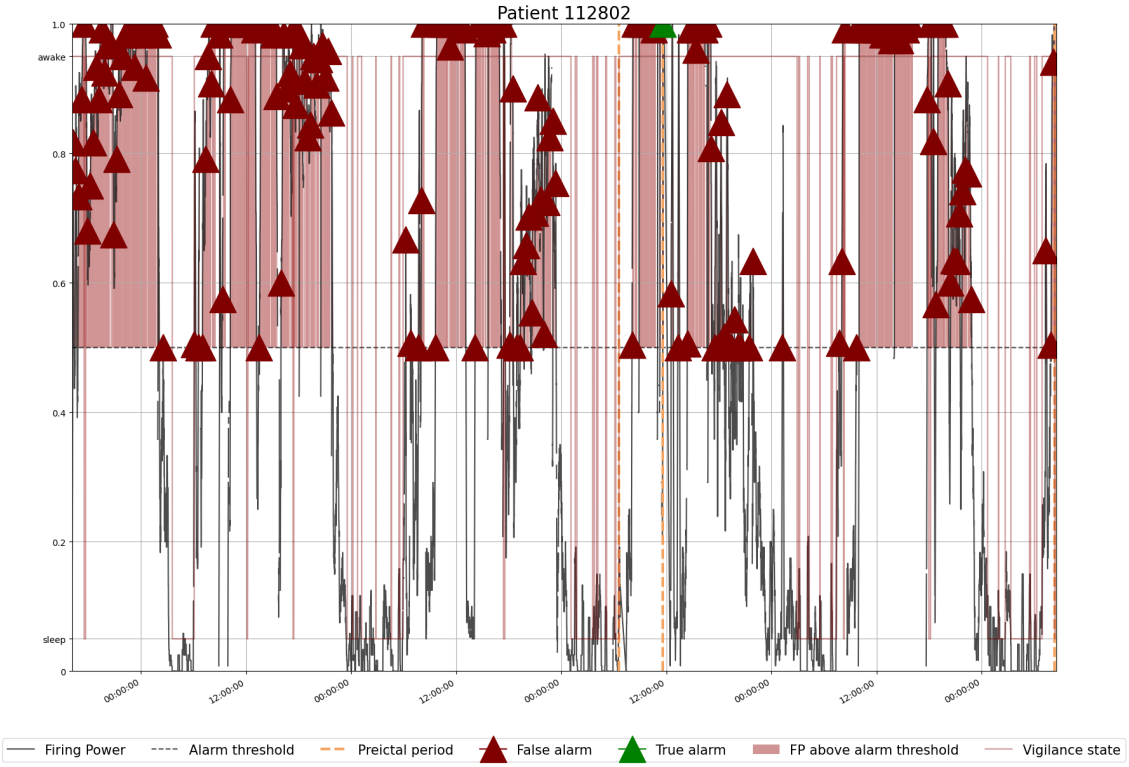

Figure 5.6: Plot of the model decisions over time for patient 8902 considering all seizure. Only the ensemble voting system is depicted in black. The dashed orange line represents the preictal period of each seizure.The vigilance state is also represented.

Furthermore, it was noticed that the three testing seizures occurred within that interval. However, only the second seizure was predicted, and the last one was not predicted due to the refractory time. Additionally, four more patients (59102, 93402, 95202, 109502) were found with similar behavior.

Wrong output

Finally, the cases in which the classifier output was entirely incorrect were also inspected. Several patients presented some seizures that were not predicted, and a few false alarms were triggered. Although in patients 98202 and 123902, it was verified that no alarm was raised, the classifier was entirely incorrect, as illustrated

in Figure 5.7.

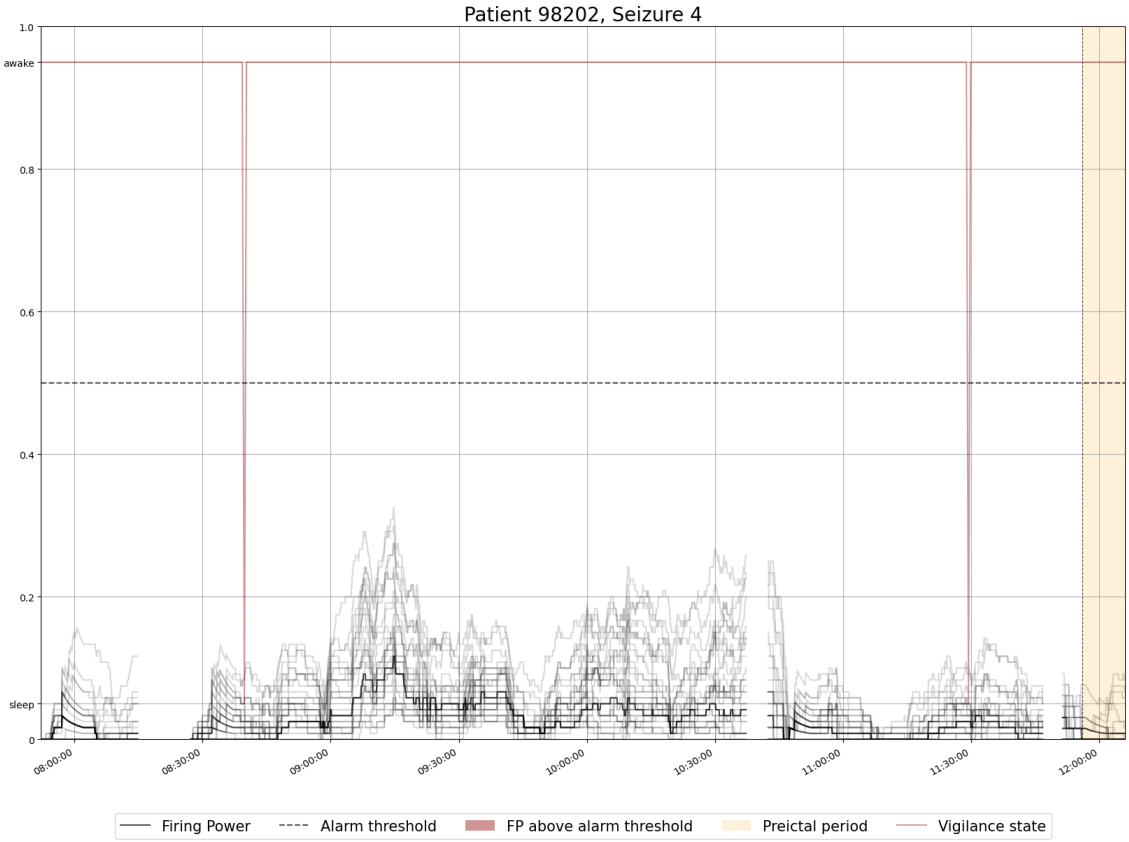

Figure 5.7: Plot of the ensemble of SVMs' decisions over time for patient 98202 seizure #4. Each SVM decision is in grey and the ensemble voting system is in black. The vigilance state is also represented.

5.2.1.1 Statistical validation

After the detailed analysis, some prominent patterns were identified, and the number of patients where a given condition is verified was counted (see Table 5.5). After that, a statistical validation was performed for each characteristic under study, as explained in Section 4.3.2.2. The parameters used in the statistical evaluation and the final results are presented in Table 5.6.

As reported in Table 5.5, 21 patients presented at least one seizure that was not predicted, but the firing power curve was trustful. On the other hand, 12 patients presented at least one seizure that was not predicted but was noted by a sleep-wake transition. For both phenomena to be statistically significant, at least nine occurrences are required since two binomial distributions were employed. In this way, it was verified that both characteristics are statistically significant and are not merely results of chance (see Table 5.6).

Table 5.5: Analysis of the presence of the identified patterns in each patient.

Patient	Seizure is not predicted but the firing power curve is trustful	Seizure detected by sleep-wake transition	Circadian cycles
402		●	
8902			
11002			
16202			
21902			
23902	●		
26102			
30802	●	●	
32702	●		
45402	●		
46702	●		
50802	●		
52302			
53402			
55202	●	●	
56402			
58602	●		
59102	●		●
60002	●		
64702		●	
75202		●	
80702	●		
85202	●	●	
93402			●
93902	●		
94402	●	●	
95202	●		●
96002		●	
98102	●		
98202		●	
101702	●		
102202	●		
104602		●	
109502			●
110602	●		
112802	●	●	●
113902			
114702			
114902	●	●	
123902			
Total	21	12	5

Regarding the circadian cycle influence, five patients presented repetitive patterns in the regularization curve (see Table 5.5). Therefore, it is possible to conclude that this characteristic is statistically significant since the number of occurrences is superior to four, the significative number of events for one binomial distribution (see Table 5.6).

Finally, considering the circadian forecasting model, the number of cases for which this model performed better and worse than the model under study was val-

idated. As mentioned in Section 5.2.1, the circadian model outperformed the SVM model in four patients, while the SVM model only outperformed the circadian model in three patients. However, none of the situations is statistically significant once the number of observations is not superior to four (see Table 5.6). Therefore, there are no conditions to raise the question related to the advantage of developing complex models using the EEG information when a simple seizure risk model that uses circadian seizure information presents a better performance for a superior number of patients. Furthermore, it would be complex and controversial to discuss since the EEG databases used in seizure prediction studies present several limitations. The EEG signals are mainly collected from patients during the pre-surgical monitoring and comprise a few hours/days of record, which is insufficient to extract definite conclusions, especially those that include circadian cycles.

Table 5.6: Parameters used in the statistical validation, the total number of occurrences of each phenomenon, and the final result (statistically valid or not).

	Seizure is not predicted but the firing power curve is trustfull	**Seizure detected by sleep-wake transition**	**Presence of circadian cycles**	**Circadian model better than SVM**	**Circadian model worse than SVM**
Binomial distribution parameters	p = 0.123 n = 40 x = 0:40			p = 0.05 n = 40 x = 0:40	
Significance level	$\alpha = 0.025$			$\alpha = 0.05$	
Statistically valid	$N > 9$ patients			$N > 4$ patients	
Total number of occurrences	21	12	5	4	3
Result	**Valid!**	**Valid!**	**Valid!**	**Invalid!**	**Invalid!**

5.2.2 Feature analysis

The features selected for each patient were evaluated. However, in general, no conclusion was obtained regarding the selected measures since, in most cases, there weren't predominant feature types, expecting for patient 8902.

As seen in Section 5.1.2, patient 8902 presented the best performance, with a sensitivity of 1 and an FPR/h of 0.16. Nevertheless, it was verified that for this patient, the algorithm selected only gamma-related features (see Figure 5.8). Regarding the electrode selection, all of them were chosen more frequently than others.

A beeswarm summary plot of SHAP Values was also produced to assess the influence of the most critical features on the model's output. The ten most significant features are presented for patient 8902 (see Figure 5.9). Notably, all shown features influence the model's behavior similarly: in general, the higher feature values have a

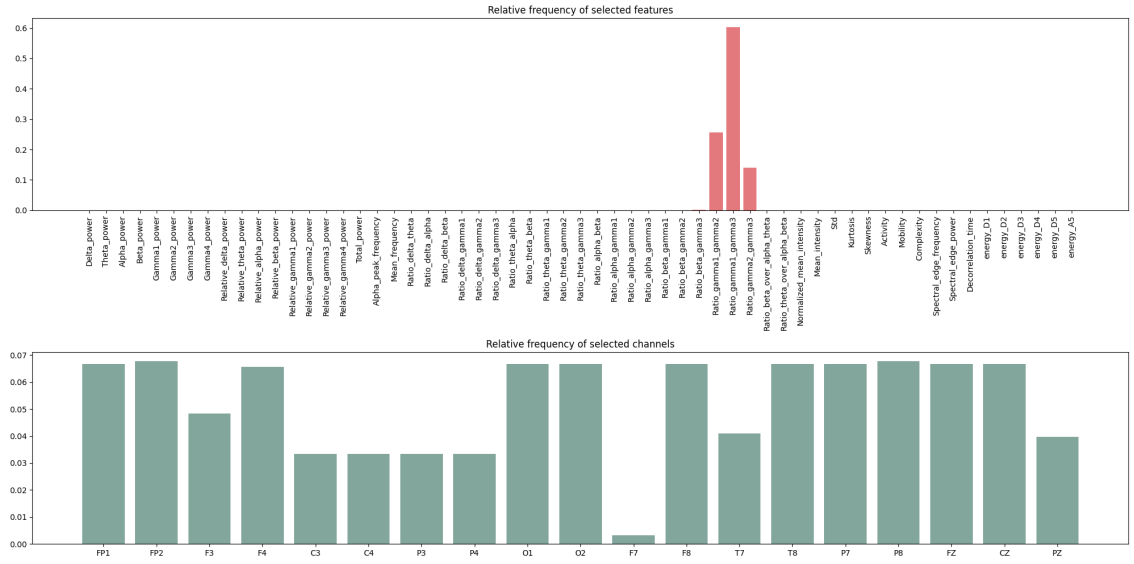

Figure 5.8: Relative frequency of the selected features and channels for patient 8902.

positive impact on the model's output, which means that when the shown features present high values, the model tends to classify the instance as preictal period. On the other hand, the lower feature values have a negative impact on the model's output, tending to the interictal classification.

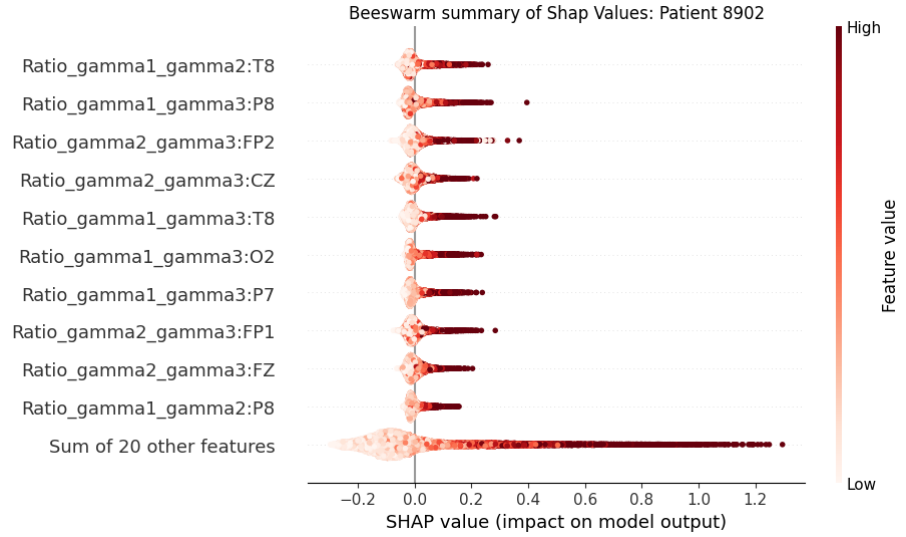

Figure 5.9: Beeswarm summary plot of Shap Values for patient 8902.

This situation may raise some skepticism once the scalp EEG does not entirely capture gamma rhythms, and thus, these features might involve muscle artifacts. Therefore, the following hypothesis can be formulated: the patient might present muscle jerks resulting from pre-seizure dynamics captured by gamma-related char-

acteristics.

The best way to corroborate this hypothesis might be using the video-EEG of the patient to confirm the muscular activity before the seizure or using iEEG since it presents a much higher signal-to-noise ratio and is less susceptible to muscular artifacts. However, this data is not available. Another form to validate this idea would be visually inspecting the EEG windows where gamma spectral-band power is significant to verify this activity. Nonetheless, it may be a time-consuming task for which an EEG specialist would be necessary to analyze several windows. It would also be possible to employ advanced tools to categorize signal epochs into artifacts, noise, or EEG-related phenomena, which would even take time. Consequently, a different strategy was applied: train new models in which the gamma-related features were discarded from the features extraction process. Furthermore, five new models were retrained to understand each spectral-band influence, using only spectral-band features from each band separately (gamma, delta, theta, beta, and alpha).

The results obtained for the new models are present in Table 5.7. It is visible that the performance worsens when the gamma-related features are eliminated. When only gamma-related features are used, the model's performance is the same, which would already be expected. It is also important to note that the optimal SOP can differ for a distinct feature set. In this case, the optimal SOP of the gamma-related model is the same as the all-feature algorithm (20 minutes), and the remaining models present an optimal SOP of 15 minutes. All the retrained models are still statistically valid, i.e. perform above chance.

By comparing each time plot (see Figure 5.10), it is possible to comprehend the model's behavior. Observing seizure #4 from patient 8902 makes it possible to verify that the model based only on gamma-related features shows a matching firing power dynamic with the pipeline under study. In contrast, the remaining models present similar relative morphology with differences in their intensities. The delta band raised more false alarms than the others.

Regarding seizure #5, it is notable that the delta and alpha bands present the most distinct behavior and more false alarms. On the other hand, the beta and theta bands and the model without gamma-related features show a similar dynamic and fewer false alarms than the all-feature and gamma-related models.

With these results, we can relieve some of the skepticism from the gamma-band features once the same patterns appear in most new models, indicating a general EEG background transition due to pre-seizure dynamics.

The same procedure was applied to three other patients to evaluate if the model's behavior presented some alterations relatively to the pipeline under study.

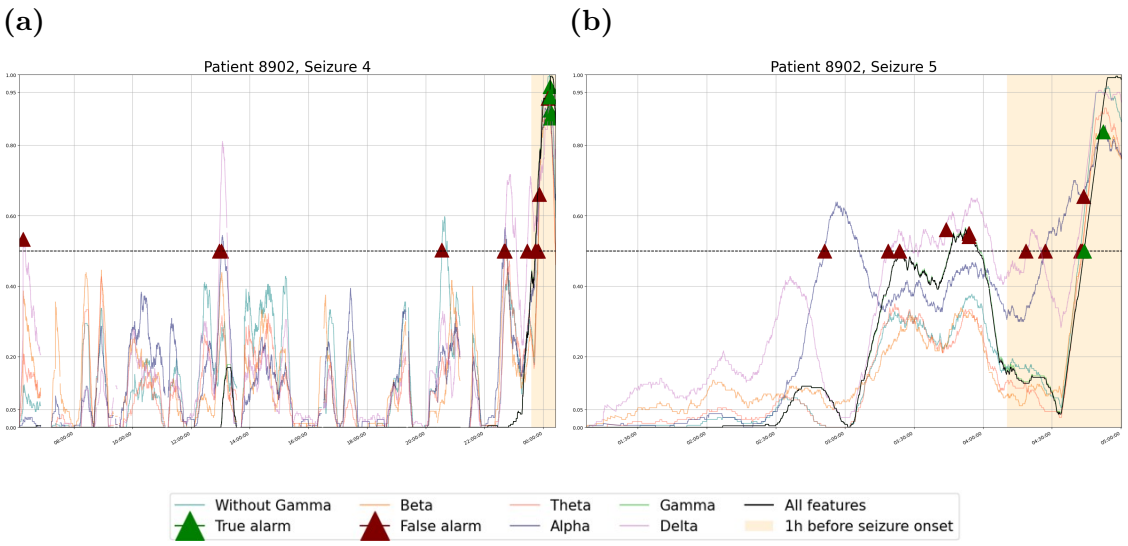

Figure 5.10: Firing power time plots of patient 8902 for different models considering distinct sets of features. The black line represents the original model, the remaining concern SVM models trained with only a determined spectral band or without gamma band. All the testing seizures are represented (**(a)**, and **(b)**).

The following patients were used:

- 94402: a case in which the prediction model behavior was inaccurate, but it was possible to observe a sleep-wake transition in the preictal period.
- 98202: a case in which the classifier output is entirely incorrect without raised alarms.
- 112802: a case in which it is possible to identify a circadian-cycle influence.

By analyzing the results for patient 94402, presented in Table 5.7, it is possible to verify that all new models, except the beta and gamma-related models, had a better performance than the algorithm under study. In these cases, the number of predicted seizures is not superior, but the number of false alarms is smaller. Being the most reduced FPR/h obtained to the approach without gamma-related features. However, even the models with better performances are still not statistically valid. It is also visible that the beta band could not predict any seizure. The optimal SOP value is only different for the theta and without gamma approaches (50 minutes) compared to the initial model (10 minutes).

The reduced FPR/h value of the without gamma model can be explained by the high SOP value (50 minutes), which leads to fewer false alarms per hour or the absence of gamma-related features. To better understand this situation, the firing power plots of both approaches were compared (see Figure 5.11).

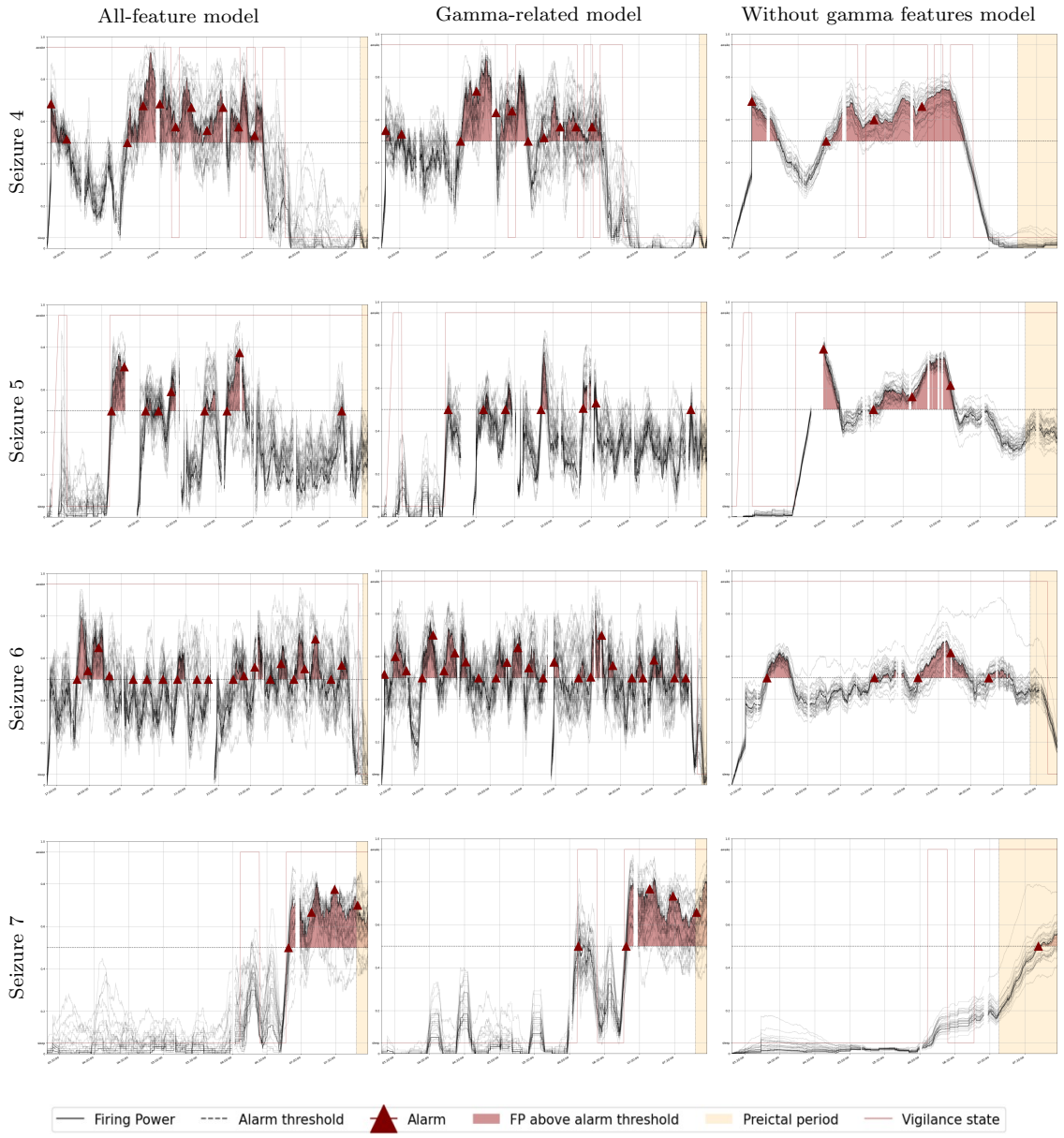

Figure 5.11: Firing power time plots of patient 94402 for different models considering distinct sets of features: all-feature model (left column), gamma-related model (middle column), and without gamma features model (right column). All the testing seizures are represented.

Notably, the model without gamma-related features presents much lesser firing power peaks superior to the threshold than the gamma model, being more evident in seizures 6 and 7. It may lead to the hypothesis that the gamma-related features may comprise muscle artifacts which influence the classifier to raise more false alarms. Indeed, it is visible that the false gamma alarms are superior during the daily period to during the night, which the presence of muscular artifacts can explain.

Furthermore, comparing the initial model (with all features) with the other two, it is evident that the first one has a firing power dynamic very similar to the gamma-related model, which shows the influence of this type of feature in the model classification. On the other hand, when the gamma-related measures are eliminated from the features set, the classifier presents an entirely different behavior with fewer firing power peaks. This analysis proves the presence of artifacts in the gamma-related features, leading to a higher FPR/h in the all-feature approach.

Finally, the firing power curve over time (see Figure 5.12) shows that, generally, the dynamic is identical through all approaches. It is worth noting that while theta, gamma, delta, alpha, and all features models predicted the seventh seizure, the model without gamma-related features predicted the sixth seizure.

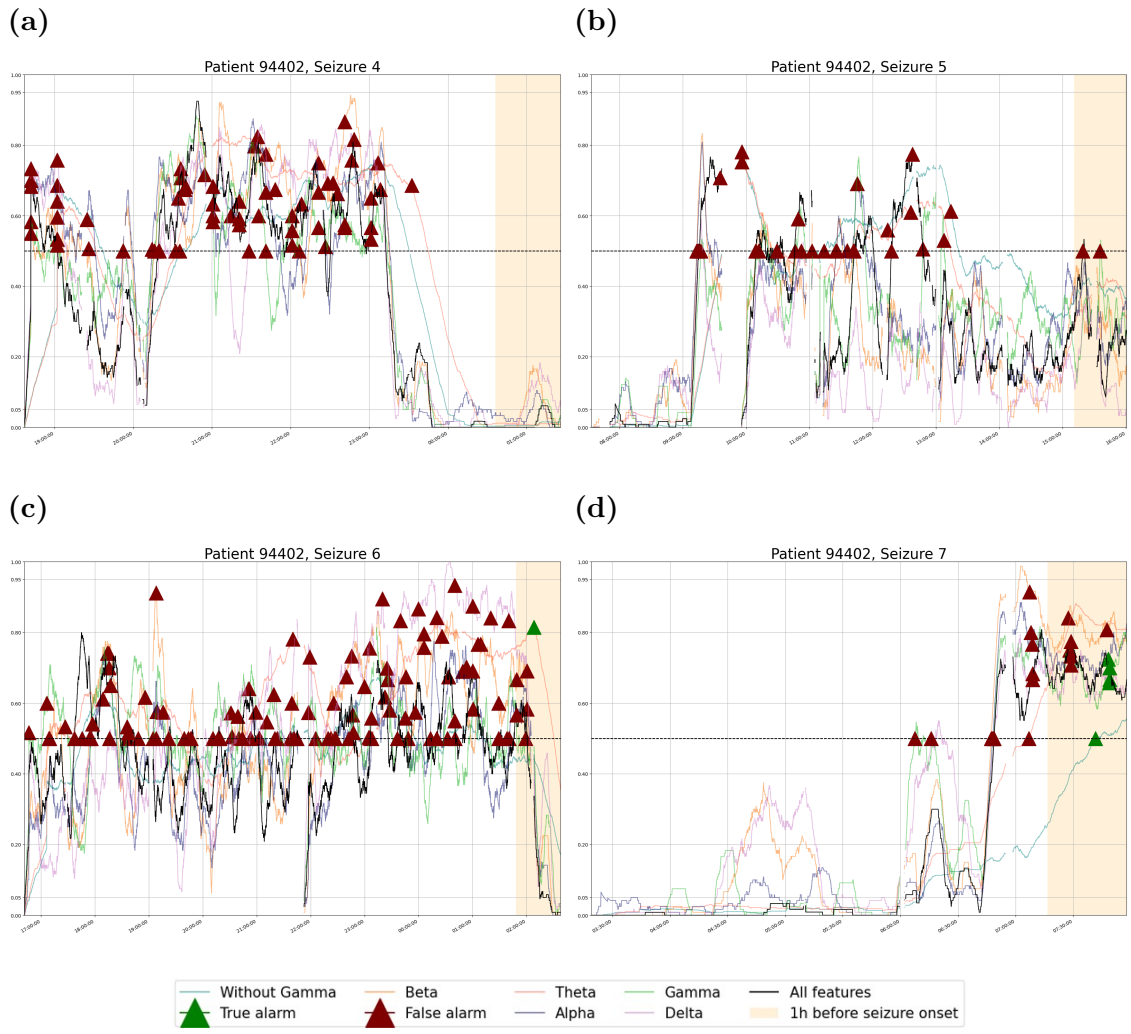

Figure 5.12: Firing power time plots of patient 94402 for different models considering distinct sets of fetaures. The black line represents the original model, the remaining concern SVM models trained with only a determined spectral band or without gamma band. All the testing seizures are represented (**(a)**, **(b)**, **(c)** , and **(d)**).

Regarding patient 98202, except for the delta band approach, all the new models present worst performance than the initial one once they raised more false alarms. The delta-related model predicted one of the seizures, raising relatively few false alarms. Its FPR/h value is 0.17, which is very close to the maximum value considered adequate (FPR/h=0.15). Furthermore, this approach is statistically valid contrary to the others. The optimal SOP's value is only different for the theta model.

Analyzing the firing power curve of all approaches (see Figure 5.13), it is notable that the gamma band produces a significant quantity of false alarms, followed by theta and beta bands that also raised several false positives.

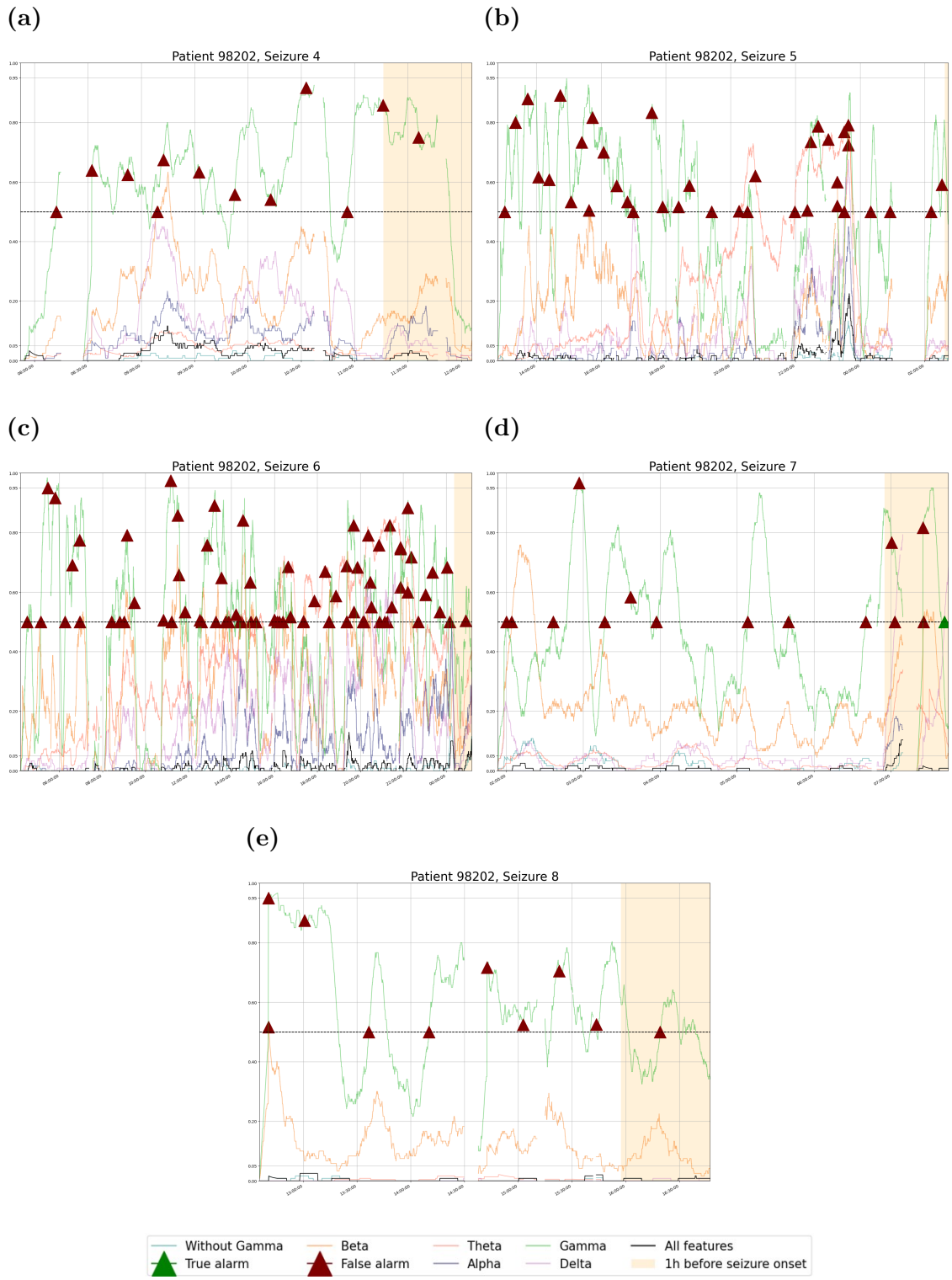

Figure 5.13: Firing power time plots of patient 98202 for different models considering distinct sets of features. The black line represents the original model, the remaining concern SVM models trained with only a determined spectral band or without gamma band. All the testing seizures are represented (**(a)**, **(b)**, **(c)** , **(d)**, and **(e)**).

It is also possible to observe that the delta-related model predicted the seventh seizure. Visually, it is notable that this patient presents an entirely different firing power dynamic when other feature sets are considered.

Finally, for patient 112802, the theta and beta-related models performed better than the initial one since they predicted more seizures and presented a lower FPR/h value. Furthermore, beta, theta and alpha band approaches are statistically significant, with the beta showing the best performance. Despite offering a much lower FPR/h value, the theta model did not predict any seizure. The optimal SOP only changes for the delta-related model (35 minutes).

Visually inspecting the firing power plots (see Figure 5.14), verifying the presence of the false alarm clusters even in the new models is possible. Therefore, the circadian cycle influence is still present independently of the approach implemented.

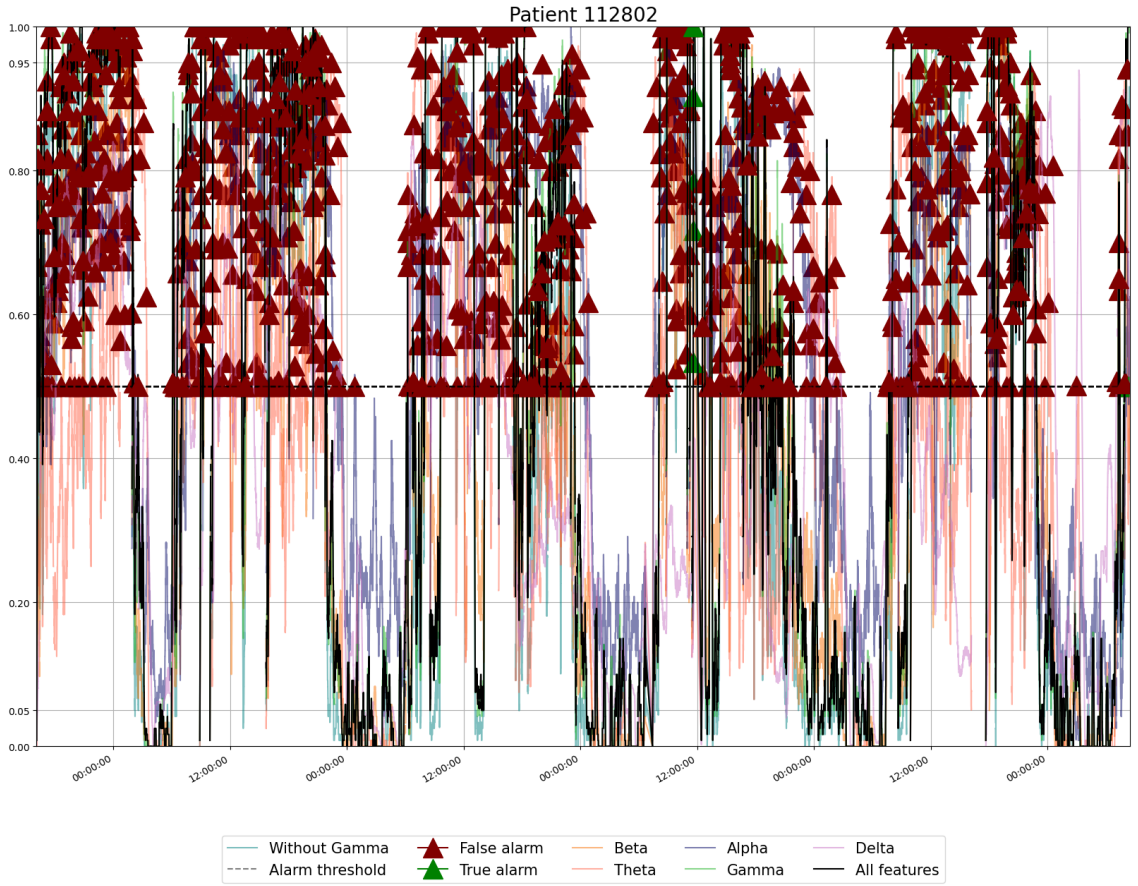

Figure 5.14: Firing power time plots of patient 112802 for different models considering distinct sets of fetaures. The black line represents the original model, the remaining concern SVM models trained with only a determined spectral band or without gamma band. All the seizures are represented.

Table 5.7: Testing performance obtained for each patient considering all approaches.

Patient	Evaluated Seizures	Approach	SOP	Predicted Seizures	False Alarms	SS	FPR/h	Statistically valid
8902	2	All features	20	2	3	1.00	0.16	●
		Alpha	15	1	6	0.50	0.33	●
		Beta	15	1	3	0.50	0.15	●
		Delta	15	1	9	0.50	0.53	●
		Gamma	20	2	3	1.00	0.16	●
		Theta	15	1	2	0.50	0.10	●
		Without Gamma	15	1	3	0.50	0.15	●
94402	4	All features	10	1	43	0.25	3.04	
		Alpha	10	1	37	0.25	2.38	
		Beta	10	0	41	0.00	2.81	
		Delta	10	1	38	0.25	2.48	
		Gamma	10	1	46	0.25	3.58	
		Theta	50	1	17	0.25	1.72	
		Without Gamma	50	1	13	0.25	1.04	
98202	5	All features	10	0	0	0.00	0.00	
		Alpha	10	0	0	0.00	0.00	
		Beta	10	0	25	0.00	0.69	
		Delta	10	1	7	0.20	0.17	●
		Gamma	10	0	98	0.00	7.09	
		Theta	20	0	23	0.00	0.71	
		Without Gamma	10	0	0	0.00	0.00	
112802	3	All features	10	1	192	0.33	4.39	
		Alpha	10	2	225	0.67	6.80	●
		Beta	10	2	175	0.67	3.62	●
		Delta	35	0	73	0.00	1.43	
		Gamma	10	1	193	0.33	4.42	
		Theta	10	2	185	0.67	4.10	●
		Without Gamma	10	1	174	0.33	3.54	

6

Conclusion

The present work aimed to explore methodologies capable of predicting epileptic seizures in ways that guarantee trust to data scientists, clinicians, and patients.

Towards this purpose, a patient-specific seizure prediction algorithm was developed based on the most common pipeline in the literature. The seizure prediction methodology achieved results of 0.34 ± 0.35 for sensitivity and 1.78 ± 1.95 for FPR/h, where 40% of patient models performed above chance.

Afterwards, different explaining strategies were employed to increase trust in the models' decisions. These explanations were based on five lessons extracted from a prior work developed by the local research team. Explaining seizure prediction models was challenging as it led to each case scenario being considered individually. Specific hypotheses were formulated and tested for each patient.

When all patients' time plots were inspected, typical model behaviors were found in a statistically significant number. This variety of patterns may represent epilepsy's clinical heterogeneity. Regarding the comparison between the circadian forecasting model, and the SVM model, it was verified that none of both models outperformed the other in a statistically significant number of patients. Finally, the impact of each spectral band over the model prediction was also evaluated, and it was concluded that, in specific scenarios, different sets of features might produce an entirely distinct behavior in the classifiers' output.

Further, patient 8902 might have been a rare case as no similar performance was achieved for any other patient. It would have been crucial to analyze more similar scenarios.

With this study, it was possible to conclude that for seizure prediction algorithms and other healthcare ML-related problems, where physiological groundings are not well established a priori, explainability should not simply explain the model's decision. It is necessary to improve the developed models, review used assumptions, and create a completer problem formulation to gain trust.

Additionally, it was notable that the evaluation of the output's classifier dy-

namic was one of the most intuitive explanations. There is the habit of merely analyzing the performance metrics of a given method. However, the present study proved that analyzing the regularization curve could provide more insights into the model's performance and decisions.

Since data from pre-surgical monitoring was used, it is noteworthy that the present study can only perform as a proof of concept. Furthermore, providing extensive examples of conjectures regarding the designed methodology is challenging once many of the developed explanation hypotheses require new testing data and extensive recording periods to capture a significant number of occurrences of a given phenomenon. For instance, despite being supported by the literature, the available data is insufficient to reinforce the hypothesized possible influence of the concept drifts' such as circadian and sleep-wake cycles.

In future work, to overcome the referred limitations, the developed methodology should be replicated in ultra-long term data collected from daily life conditions, such as those performed by Cook et al. [24]. By employing these data, finding other case scenarios and validating the suggested ones might be possible. Moreover, a long-term analysis will supply a definitive evaluation of these explanation methods as they must prevail effective and intuitive when examining days and months of records. Interviewing patients must also be considered to understand their perspectives and relation to devices and guarantee trust.

Bibliography

[1] I. E. Scheffer, S. Berkovic, G. Capovilla, M. B. Connolly, J. French, L. Guilhoto, E. Hirsch, S. Jain, G. W. Mathern, S. L. Moshé, *et al.*, "Ilae classification of the epilepsies: position paper of the ilae commission for classification and terminology," *Epilepsia*, vol. 58, no. 4, pp. 512–521, 2017.

[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.

[3] A. Varsavsky, I. Mareels, and M. Cook, *Epileptic seizures and the EEG: measurement, models, detection and prediction*. Taylor & Francis, 2011.

[4] A. K. Health, "Invasive electroencephalography (eeg) monitoring before epilepsy surgery," 2017. Available: `https://www.aboutkidshealth.ca/article?contentid=2056&language=english`. Last accessed: 29 January 2022.

[5] N. Moghim and D. W. Corne, "Predicting epileptic seizures in advance," *PloS one*, vol. 9, no. 6, p. e99334, 2014.

[6] M. Winterhalder, T. Maiwald, H. Voss, R. Aschenbrenner-Scheibe, J. Timmer, and A. Schulze-Bonhage, "The seizure prediction characteristic: a general framework to assess and compare seizure prediction methods," *Epilepsy & Behavior*, vol. 4, no. 3, pp. 318–325, 2003.

[7] B. Schelter, R. G. Andrzejak, and F. Mormann, "Can your prediction algorithm beat a random predictor?," *Seizure prediction in epilepsy: from basic mechanisms to clinical applications*, pp. 237–248, 2008.

[8] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A

review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2021.

[9] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[10] C. Molnar, *Interpretable Machine Learning*. 2019.

[11] K. Rasheed, A. Qayyum, J. Qadir, S. Sivathamboo, P. Kwan, L. Kuhlmann, T. O'Brien, and A. Razi, "Machine learning for predicting epileptic seizures using eeg signals: A review," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 139–155, 2020.

[12] W. H. Organization *et al.*, *Epilepsy: a public health imperative*. World Health Organization, 2019.

[13] O. Devinsky, A. Vezzani, T. O'Brien, *et al.*, "Epilepsy," *Nat Rev Dis Primers 4*, p. 18024, 2018.

[14] J. Engel, "What can we do for people with drug-resistant epilepsy?: The 2016 wartenberg lecture," *Neurology*, vol. 87, no. 23, pp. 2483–2489, 2016.

[15] C. Rathore and K. Radhakrishnan, "Concept of epilepsy surgery and presurgical evaluation," *Epileptic disorders*, vol. 17, no. 1, pp. 19–31, 2015.

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

[17] E. B. Assi, D. K. Nguyen, S. Rihana, and M. Sawan, "Towards accurate prediction of epileptic seizures: A review," *Biomedical Signal Processing and Control*, vol. 34, pp. 144–157, 2017.

[18] F. Mormann, R. G. Andrzejak, C. E. Elger, and K. Lehnertz, "Seizure prediction: the long and winding road," *Brain*, vol. 130, no. 2, pp. 314–333, 2007.

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

[20] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen, "Dynamic integration of classifiers for handling concept drift," *Information fusion*, vol. 9, no. 1, pp. 56–68, 2008.

[21] T. R. Hoens, R. Polikar, and N. V. Chawla, "Learning from streaming data

with concept drift and imbalance: an overview," *Progress in Artificial Intelligence*, vol. 1, no. 1, pp. 89–101, 2012.

[22] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89, IEEE, 2018.

[23] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

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

[25] F. T. Sun and M. J. Morrell, "The rns system: responsive cortical stimulation for the treatment of refractory partial epilepsy," *Expert review of medical devices*, vol. 11, no. 6, pp. 563–572, 2014.

[26] R. S. Fisher, W. V. E. Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel Jr, "Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe)," *Epilepsia*, vol. 46, no. 4, pp. 470–472, 2005.

[27] 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.

[28] E. Foundation, "Seizure onset information." Available: `https://epilepsynewengland.org/knowledge-center/types-of-seizures/seizure-onset`. Last accessed: 24 January 2022.

[29] E. Foundation, "Types of epilepsy syndromes." Availabel: `https://www.epilepsy.com/learn/types-epilepsy-syndromes`. Last accessed: 25 January 2022.

[30] E. Foundation, "Temporal lobe epilepsy (tle)." Available: `https://www.epilepsy.com/learn/types-epilepsy-syndromes/temporal-lobe-epilepsy-aka-tle`. Last accessed: 25 January 2022.

[31] N. I. of Neurological Disorders and S. (NIHDS), "The epilepsies and seizures: Hope through research." Available: https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Hope-Through-Research/Epilepsies-and-Seizures-Hope-Through. Last accessed: 25 January 2022.

[32] E. Foundation, "Acute repetitive seizures (ars) or cluster seizures." Available: https://www.epilepsyfoundationmn.org/2020/01/14/acute-repetitive-seizures-ars-or-cluster-seizures/. Last accessed: 18 May 2022.

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

[34] I. Osorio, H. P. Zaveri, M. G. Frei, and S. Arthurs, *Epilepsy: the intersection of neurosciences, biology, mathematics, engineering, and physics*. CRC press, 2019.

[35] H. Chen and M. Z. Koubeissi, "Electroencephalography in epilepsy evaluation," *CONTINUUM: Lifelong Learning in Neurology*, vol. 25, no. 2, pp. 431–453, 2019.

[36] T. N. Alotaiby, S. A. Alshebeili, T. Alshawi, I. Ahmad, and F. E. Abd El-Samie, "Eeg seizure detection and prediction algorithms: a survey," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1–21, 2014.

[37] J. W. C. Medithe and U. R. Nelakuditi, "Study of normal and abnormal eeg," in *2016 3rd International conference on advanced computing and communication systems (ICACCS)*, vol. 1, pp. 1–4, IEEE, 2016.

[38] F. Amzica, "What does burst suppression really mean?," *Epilepsy & Behavior*, vol. 49, pp. 234–237, 2015.

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

[40] E. Foundation, "Video eeg monitoring with invasive electrodes." Available: https://www.epilepsy.com/learn/treating-seizures-and-epilepsy/surgery/tests-surgery/video-eeg-monitoring-invasive-electrodes. Last accessed: 29 January 2022.

[41] S. Ramgopal, S. Thome-Souza, M. Jackson, N. E. Kadish, I. S. Fernández,

J. Klehm, W. Bosl, C. Reinsberger, S. Schachter, and T. Loddenkemper, "Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy," *Epilepsy & behavior*, vol. 37, pp. 291–307, 2014.

[42] S. B. Dumanis, J. A. French, C. Bernard, G. A. Worrell, and B. E. Fureman, "Seizure forecasting from idea to reality. outcomes of the my seizure gauge epilepsy innovation institute workshop," *Eneuro*, vol. 4, no. 6, 2017.

[43] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.

[44] B. Schelter, M. Winterhalder, T. Maiwald, A. Brandt, A. Schad, A. Schulze-Bonhage, and J. Timmer, "Testing statistical significance of multivariate time series analysis techniques for epileptic seizure prediction," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 16, no. 1, p. 013108, 2006.

[45] R. G. Andrzejak, F. Mormann, T. Kreuz, C. Rieke, A. Kraskov, C. E. Elger, and K. Lehnertz, "Testing the null hypothesis of the nonexistence of a preseizure state," *Physical Review E*, vol. 67, no. 1, p. 010901, 2003.

[46] J. Klatt, H. Feldwisch-Drentrup, M. Ihle, V. Navarro, M. Neufang, C. Teixeira, C. Adam, M. Valderrama, C. Alvarado-Rojas, A. Witon, *et al.*, "The epilepsiae database: An extensive electroencephalography database of epilepsy patients," 2012.

[47] R. Elshawi, M. H. Al-Mallah, and S. Sakr, "On the interpretability of machine learning-based model for predicting hypertension," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–32, 2019.

[48] L. H. Gilpin, C. Testart, N. Fruchter, and J. Adebayo, "Explaining explanations to society," *arXiv preprint arXiv:1901.06560*, 2019.

[49] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[50] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *Advances in neural information processing systems*, vol. 29, 2016.

[51] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou, "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond," *arXiv preprint arXiv:2103.10689*, 2021.

[52] T. J. Kennedy and L. A. Lingard, "Making sense of grounded theory in medical education," *Medical education*, vol. 40, no. 2, pp. 101–108, 2006.

[53] H. Engward, "Understanding grounded theory," *Nursing Standard (through 2013)*, vol. 28, no. 7, p. 37, 2013.

[54] W. Miller, "Understanding grounded theory," *Clinical Laboratory Science*, vol. 28, no. 3, pp. 197–200, 2015.

[55] A. H. Thomas, A. Aminifar, and D. Atienza, "Noise-resilient and interpretable epileptic seizure detection," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, IEEE, 2020.

[56] C. Teixeira, B. Direito, M. Bandarabadi, and A. Dourado, "Output regularization of svm seizure predictors: Kalman filter versus the "firing power" method," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6530–6533, IEEE, 2012.

[57] M. Bandarabadi, C. A. Teixeira, B. Direito, and A. Dourado, "Epileptic seizure prediction based on a bivariate spectral power methodology," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5943–5946, IEEE, 2012.

[58] J. Rasekhi, M. R. K. Mollaei, M. Bandarabadi, C. A. Teixeira, and A. Dourado, "Preprocessing effects of 22 linear univariate features on the performance of seizure prediction methods," *Journal of neuroscience methods*, vol. 217, no. 1-2, pp. 9–16, 2013.

[59] A. F. Rabbi, L. Azinfar, and R. Fazel-Rezai, "Seizure prediction using adaptive neuro-fuzzy inference system," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2100–2103, IEEE, 2013.

[60] C. Alvarado-Rojas, M. Valderrama, A. Fouad-Ahmed, H. Feldwisch-Drentrup, M. Ihle, C. Teixeira, F. Sales, A. Schulze-Bonhage, C. Adam, A. Dourado, *et al.*, "Slow modulations of high-frequency activity (40–140 hz) discriminate preictal changes in human focal epilepsy," *Scientific reports*, vol. 4, no. 1, pp. 1–9, 2014.

[61] C. A. Teixeira, B. Direito, M. Bandarabadi, M. Le Van Quyen, M. Valderrama, B. Schelter, A. Schulze-Bonhage, V. Navarro, F. Sales, and A. Dourado, "Epileptic seizure predictors based on computational intelligence techniques:

A comparative study with 278 patients," *Computer methods and programs in biomedicine*, vol. 114, no. 3, pp. 324–336, 2014.

[62] J. Rasekhi, M. R. K. Mollaei, M. Bandarabadi, C. A. Teixeira, and A. Dourado, "Epileptic seizure prediction based on ratio and differential linear univariate features," *Journal of medical signals and sensors*, vol. 5, no. 1, p. 1, 2015.

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

[64] S. M. Usman, M. Usman, and S. Fong, "Epileptic seizures prediction using machine learning methods," *Computational and mathematical methods in medicine*, vol. 2017, 2017.

[65] A. Aarabi and B. He, "Seizure prediction in patients with focal hippocampal epilepsy," *Clinical Neurophysiology*, vol. 128, no. 7, pp. 1299–1307, 2017.

[66] B. Direito, C. A. Teixeira, F. Sales, M. Castelo-Branco, and A. Dourado, "A realistic seizure prediction study based on multiclass svm," *International journal of neural systems*, vol. 27, no. 03, p. 1750006, 2017.

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

[68] I. Kiral-Kornek, S. Roy, E. Nurse, B. Mashford, P. Karoly, T. Carroll, D. Payne, S. Saha, S. Baldassano, T. O'Brien, *et al.*, "Epileptic seizure prediction using big data and deep learning: toward a mobile system," *EBioMedicine*, vol. 27, pp. 103–111, 2018.

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

[70] S. M. Usman and A. Hassan, "Efficient prediction and classification of epileptic seizures using eeg data based on univariate linear features.," *J. Comput.*, vol. 13, no. 6, pp. 616–621, 2018.

[71] L. A. S. Kitano, M. A. A. Sousa, S. D. Santos, R. Pires, S. Thome-Souza, and A. B. Campo, "Epileptic seizure prediction from eeg signals using unsupervised

learning and a polling-based decision process," in *International Conference on Artificial Neural Networks*, pp. 117–126, Springer, 2018.

[72] S. Yuan, W. Zhou, and L. Chen, "Epileptic seizure prediction using diffusion distance and bayesian linear discriminate analysis on intracranial eeg," *International Journal of Neural Systems*, vol. 28, no. 01, p. 1750043, 2018.

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

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

[75] N. D. Truong, L. Kuhlmann, M. R. Bonyadi, D. Querlioz, L. Zhou, and O. Kavehei, "Epileptic seizure forecasting with generative adversarial networks," *IEEE Access*, vol. 7, pp. 143999–144009, 2019.

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

[77] A. Gabara, R. Yousri, D. Hamdy, M. H. Zakhari, and H. Mostafa, "Patient specific epileptic seizures prediction based on support vector machine," in *2020 32nd International Conference on Microelectronics (ICM)*, pp. 1–4, IEEE, 2020.

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

[79] T. Tamanna, M. A. Rahman, S. Sultana, M. H. Haque, and M. Z. Parvez, "Predicting seizure onset based on time-frequency analysis of eeg signals," *Chaos, Solitons & Fractals*, vol. 145, p. 110796, 2021.

[80] M. Pinto, A. Leal, F. Lopes, A. Dourado, P. Martins, C. A. Teixeira, *et al.*, "A personalized and evolutionary algorithm for interpretable eeg epilepsy seizure prediction," *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.

[81] S. M. Usman, S. Khalid, and Z. Bashir, "Epileptic seizure prediction using scalp electroencephalogram signals," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 1, pp. 211–220, 2021.

[82] P. Peng, Y. Song, and L. Yang, "Seizure prediction in eeg signals using stft and domain adaptation," *Frontiers in Neuroscience*, p. 1880, 2021.

[83] K. Singh and J. Malhotra, "Predicting epileptic seizures from eeg spectral band features using convolutional neural network," *Wireless Personal Communications*, pp. 1–18, 2022.

[84] D. Liang, A. Liu, C. Li, J. Liu, and X. Chen, "A novel consistency-based training strategy for seizure prediction," *Journal of Neuroscience Methods*, vol. 372, p. 109557, 2022.

[85] M. Pinto, T. Coelho, A. Leal, F. Lopes, A. Dourado, P. Martins, and C. Teixeira, "Interpretable eeg seizure prediction using a multiobjective evolutionary algorithm," *Scientific reports*, vol. 12, no. 1, pp. 1–15, 2022.

[86] T. Pal Attia, P. F. Viana, M. Nasseri, J. Duun-Henriksen, A. Biondi, J. S. Winston, I. P. Martins, E. S. Nurse, M. Dümpelmann, G. A. Worrell, *et al.*, "Seizure forecasting using minimally invasive, ultra-long-term subcutaneous eeg: Generalizable cross-patient models," *Epilepsia*, 2022.

[87] M. Nasseri, T. Pal Attia, B. Joseph, N. M. Gregg, E. S. Nurse, P. F. Viana, G. Worrell, M. Dümpelmann, M. P. Richardson, D. R. Freestone, *et al.*, "Ambulatory seizure forecasting with a wrist-worn device using long-short term memory deep learning," *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[88] B. Maimaiti, H. Meng, Y. Lv, J. Qiu, Z. Zhu, Y. Xie, Y. Li, W. Zhao, J. Liu, M. Li, *et al.*, "An overview of eeg-based machine learning methods in seizure prediction and opportunities for neurologists in this field," *Neuroscience*, 2021.

[89] S. Serrano and N. A. Smith, "Is attention interpretable?," *arXiv preprint arXiv:1906.03731*, 2019.

[90] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[91] D. Barić, P. Fumić, D. Horvatić, and T. Lipic, "Benchmarking attention-based interpretability of deep learning in multivariate time series predictions," *Entropy*, vol. 23, no. 2, p. 143, 2021.

[92] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Interpretable seizure classification using unprocessed eeg with multi-channel attentive feature fusion," *IEEE Sensors Journal*, vol. 21, no. 17, pp. 19186–19197, 2021.

[93] A. Baghdadi, S. Daoud, M. Dammak, C. Mhiri, P. Siarry, A. M. Alimi, *et al.*, "A channel-wise attention-based representation learning method for epileptic seizure detection and type classification," 2021.

[94] X. Zhang, L. Yao, M. Dong, Z. Liu, Y. Zhang, and Y. Li, "Adversarial representation learning for robust patient-independent epileptic seizure detection," *IEEE journal of biomedical and health informatics*, vol. 24, no. 10, pp. 2852–2859, 2020.

[95] T.-Y. Hsieh, S. Wang, Y. Sun, and V. Honavar, "Explainable multivariate time series classification: A deep neural network which learns to attend to important variables as well as time intervals," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 607–615, 2021.

[96] M. Mansour, F. Khnaisser, and H. Partamian, "An explainable model for eeg seizure detection based on connectivity features," *arXiv preprint arXiv:2009.12566*, 2020.

[97] V. Gabeff, T. Teijeiro, M. Zapater, L. Cammoun, S. Rheims, P. Ryvlin, and D. Atienza, "Interpreting deep learning models for epileptic seizure detection on eeg signals," *Artificial Intelligence in Medicine*, vol. 117, p. 102084, 2021.

[98] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for patient-independent epileptic seizure prediction using scalp eeg signals," *IEEE Sensors Journal*, vol. 21, no. 7, pp. 9377–9388, 2021.

[99] G. Wang, Z. Deng, and K.-S. Choi, "Detection of epilepsy with electroencephalogram using rule-based classifiers," *Neurocomputing*, vol. 228, pp. 283–290, 2017.

[100] M. Lo Giudice, G. Varone, C. Ieracitano, N. Mammone, G. G. Tripodi, E. Ferlazzo, S. Gasparini, U. Aguglia, and F. C. Morabito, "Permutation entropy-based interpretability of convolutional neural network models for interictal eeg discrimination of subjects with epileptic seizures vs. psychogenic non-epileptic seizures," *Entropy*, vol. 24, no. 1, p. 102, 2022.

[101] S. Naze, J. Tang, J. R. Kozloski, and S. Harrer, "Features importance in seizure classification using scalp eeg reduced to single timeseries," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 329–332, IEEE, 2021.

[102] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Geo-

metric deep learning for subject-independent epileptic seizure prediction using scalp eeg signals," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[103] S. Tang, J. Dunnmon, K. K. Saab, X. Zhang, Q. Huang, F. Dubost, D. Rubin, and C. Lee-Messer, "Self-supervised graph neural networks for improved electroencephalographic seizure analysis," in *International Conference on Learning Representations*, 2021.

[104] T. Uyttenhove, A. Maes, T. Van Steenkiste, D. Deschrijver, and T. Dhaene, "Interpretable epilepsy detection in routine, interictal eeg data using deep learning," in *Machine Learning for Health*, pp. 355–366, PMLR, 2020.

[105] M. S. Hossain, S. U. Amin, M. Alsulaiman, and G. Muhammad, "Applying deep learning for epilepsy seizure detection and brain mapping visualization," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1s, pp. 1–17, 2019.

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

[107] D. E. Moghaddam, S. Sheth, Z. Haneef, J. Gavvala, and B. Aazhang, "Epileptic seizure prediction using spectral width of the covariance matrix," *Journal of Neural Engineering*, 2022.

[108] C. A. Ellis, R. L. Miller, and V. D. Calhoun, "A novel local explainability approach for spectral insight into raw eeg-based deep learning classifiers," in *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 1–6, IEEE, 2021.

[109] M. Briden and N. Norouzi, "Wavefusion squeeze-and-excitation: Towards an accurate and explainable deep learning framework in neuroscience," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1092–1095, IEEE, 2021.

[110] F. Andreotti, H. Phan, and M. De Vos, "Visualising convolutional neural network decisions in automatic sleep scoring," in *CEUR Workshop Proceedings*, pp. 70–81, CEUR Workshop Proceedings, 2018.

[111] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "Salientsleepnet: Multimodal salient wave detection network for sleep staging," *arXiv preprint arXiv:2105.13864*, 2021.

[112] H. Phan, K. B. Mikkelsen, O. Chen, P. Koch, A. Mertins, and M. De Vos, "Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Transactions on Biomedical Engineering*, 2022.

[113] A. Vilamala, K. H. Madsen, and L. K. Hansen, "Deep convolutional neural networks for interpretable analysis of eeg sleep stage scoring," in *2017 IEEE 27th international workshop on machine learning for signal processing (MLSP)*, pp. 1–6, IEEE, 2017.

[114] I. Al-Hussaini, C. Xiao, M. B. Westover, and J. Sun, "Sleeper: interpretable sleep staging via prototypes from expert rules," in *Machine Learning for Healthcare Conference*, pp. 721–739, PMLR, 2019.

[115] F. Lopes, A. Leal, J. Medeiros, M. F. Pinto, A. Dourado, M. Dümpelmann, and C. Teixeira, "Automatic electroencephalogram artifact removal using deep convolutional neural networks," *IEEE Access*, vol. 9, pp. 149955–149970, 2021.

[116] Y. Yoo, "On predicting epileptic seizures from intracranial electroencephalography," *Biomedical engineering letters*, vol. 7, no. 1, pp. 1–5, 2017.

[117] A. S. Malik and H. U. Amin, *Designing EEG experiments for studying the brain: Design code and example datasets*. Academic Press, 2017.

[118] C. Herrmann and T. Demiralp, "Human eeg gamma oscillations in neuropsychiatric disorders," *Clinical neurophysiology*, vol. 116, no. 12, pp. 2719–2733, 2005.

[119] A. V. Medvedev, "Epileptiform spikes desynchronize and diminish fast (gamma) activity of the brain: an "anti-binding" mechanism?," *Brain research bulletin*, vol. 58, no. 1, pp. 115–128, 2002.

[120] K. Kobayashi, M. Oka, T. Akiyama, T. Inoue, K. Abiru, T. Ogino, H. Yoshinaga, Y. Ohtsuka, and E. Oka, "Very fast rhythmic activity on scalp eeg associated with epileptic spasms," *Epilepsia*, vol. 45, no. 5, pp. 488–496, 2004.

[121] T. R. C. H. Melbourne, "Rectal diazepam." Available: `https://www.rch.org.au/neurology/patient_information/rectal_diazepam/`. Last accessed: 20 August 2022.

[122] A. Schulze-Bonhage, F. Sales, K. Wagner, R. Teotonio, A. Carius, A. Schelle, and M. Ihle, "Views of patients with epilepsy on seizure prediction devices," *Epilepsy & behavior*, vol. 18, no. 4, pp. 388–396, 2010.

[123] S. Psychology, "Karl popper - theory of falsification," 2020. Last accessed 17 August 2022.

[124] S. Taran, N. K. Adhikari, and E. Fan, "Falsifiability in medicine: What clinicians can learn from karl popper," *Intensive Care Medicine*, vol. 47, no. 9, pp. 1054–1056, 2021.

[125] E. Shahar, "A popperian perspective of the term 'evidence-based medicine'," *Journal of Evaluation in Clinical practice*, vol. 3, no. 2, pp. 109–116, 1997.

[126] F. Mormann, T. Kreuz, C. Rieke, R. G. Andrzejak, A. Kraskov, P. David, C. E. Elger, and K. Lehnertz, "On the predictability of epileptic seizures," *Clinical neurophysiology*, vol. 116, no. 3, pp. 569–587, 2005.

Appendices

A

Features Description

Here are present some details regarding the extracted features. Linear features are mathematical measures that capture linear dynamics from the signal, using its phase/frequency and amplitude information. The EEG signal is assumed as quasi-stationary within each time window when this type of feature is extracted.

A.1 Statistical Moments

Statistical moments are widely used in seizure prediction studies to characterize the signal's amplitude distribution. The four moments are mean, variance, skewness, which measures the degree of asymmetries of the amplitude distribution, and kurtosis, which measures the relative flatness or peakedness of the amplitude distribution. The preictal period has been associated with considerable changes in these measures compared to the interictal period. In particular, a decrease in variance and an increase in kurtosis were observed in the preictal phase [18, 61, 65, 66, 126].

A.2 Hjörth Parameters

The Hjörth parameters consider standard deviations to quantify the dynamical signal properties. These are three time-domain measures of brain activity: activity, a measure of mean power, mobility, a measure of root-mean-squared frequency, and complexity, a measure of root-mean-square frequency spread. With the proximity to the seizure onset, an increase in mobility and complexity measures is observed [18, 58, 61, 66, 126].

A.3 Decorrelation Time

The decorrelation time is described as the first zero crossing of the autocorrelation function. It is an estimator of the data periodicity and the strength of linear

correlations. The lower its values, the less the signal is correlated. Before seizures, a decrease in the decorrelation time has been reported [61, 66].

A.4 Relative Spectral Power

The spectral power quantifies the signal power associated with specific frequency ranges. It is possible to compute the power spectral density (PSD) by applying the Fast Fourier Transform (FFT) to the EEG time series and then average the squared coefficients of the frequency range of interest.

In turn, the relative spectral power is characterized as the power of a given frequency band divided by the total power of the EEG signal. A normalized spectral power provides a more robust measure since there is more power in low frequencies than at high frequencies. Some authors have reported a transference of power from the lower to higher frequencies before the seizure onset [58, 61, 66, 126].

A.5 Spectral Edge Frequency and Power

Spectral Edge Frequency (SEF) is commonly described as the minimum frequency below which a given percentage of the total power of the signal is contained. The Spectral Edge Power (SEP) is the value of the power existing below the defined threshold.

Regarding the EEG signal, most of the spectral power is comprised in the 0.5–40Hz band, and SEF 50 and SEP 50 are commonly used. SEF 50 is the frequency below which 50% of the total power of the signal up to 40 Hz is located, and SEP 50 is the corresponding power below the spectral edge frequency. Thus, SEF may be capable of capturing the dynamics mentioned above during the preictal [61, 66].

A.6 Wavelet Coefficients Energy

The Discrete Wavelet Transform (DWT) is a time-frequency domain transform that can be an alternative to the Fast Fourier Transform (FFT). It is capable of revealing the spectral and temporal properties of the signal. The wavelet transform decomposes the signal in different resolution levels according to specific frequency components. The first decomposition levels are associated with higher frequencies, while the last levels represent the lower frequencies. After the signal decomposition, it is possible to compute discriminant measures from distinct frequency bands by applying the wavelet coefficients. The quantification of the energy in different fre-

quency ranges is an example of a feature that can be obtained using the wavelet transform [61, 66].