UNIVERSIDADE D
COIMBRA

Mauro Filipe da Silva Pinto

TOWARDS THE INTERPRETATION OF MACHINE LEARNING SEIZURE PREDICTION MODELS

Doctoral thesis submitted in partial fulfilment of the Doctoral Program in Informatics Engineering, Intelligent Systems supervised by Professor César Alexandre Domingues Teixeira and Professor Pedro José Mendes Martins, and presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra

September 2022

Faculty of Sciences and Technology

University of Coimbra

Towards the Interpretation of Machine Learning Seizure Prediction Models

Mauro Filipe da Silva Pinto

Supervisor:
César Alexandre Domingues Teixeira

Co-supervisor:
Pedro José Mendes Martins

Dissertation presented to obtain a Ph.D. degree in Informatics Engineering, Intelligent Systems at the Faculty of Sciences and Technology of the University of Coimbra

Dissertação de Doutoramento apresentada à Faculdade de Ciências e Tecnologia da Universidade de Coimbra, para prestação de provas de Doutoramento em Engenharia Informática, Sistemas Inteligentes

September 2022

The studies presented in this thesis were carried out at the Centre for Informatics and Systems of the University of Coimbra (CISUC) at the Informatics Department, Faculty of Sciences and Technology, University of Coimbra, Portugal.

This work was supported by the Portuguese Foundation for Science and Technology (FCT), Human Capital Operational Program (POCH) and the European Union (EU) under the Ph.D. grant SFRH/BD/139757/2018, along with projects CISUC (UID/CEC/00326/2020) and project RECoD (PTDC/EEI-EEE/5788/2020), both financed by national funds through the FCT – Foundation for Science and Technology, I.P..

For my parents

É tão difícil guardar um rio
quando corre dentro de nós
Jorge Sousa Braga

Acknowledgements

I thank my supervisors, Professor César Teixeira and Professor Pedro Martins, and also Professor António Dourado for all the guidance. I thank all the professors and colleagues from the Department of Informatics Engineering that helped me along the way.

I thank the Centre of Informatics and Systems from the University of Coimbra (CISUC) for hosting me for the last four years and Fundação de Ciência e Tecnologia (FCT) for funding my work with a PhD grant.

I thank Centro Hospitalar e Universitário de Coimbra (CHUC), namely the refractory epilepsy centre, the sleep monitoring unit, and Dr Francisco Sales, for all the help.

I thank Adriana Leal and Fábio Lopes for all their friendship during the PhD and their research. I also thank the remaining colleagues and friends of our laboratory.

I thank all my students for what I have learned from them.

Last but not least, I thank my parents, Tânia, and André, for everything.

Abstract

Seizure prediction concerns a multidimensional time-series problem that typically performs continuous sliding window analysis and classification. Current state-of-the-art methods using the electroencephalogram signal are based on Machine Learning (ML) models that are mainly black boxes, weakening the trust of clinicians for high-risk decisions. Despite decades of research, few devices/systems underwent clinical trials and/or are commercialised, where these do not use the most recent approaches, such as neural networks, to their full potential. The absence of explanations for black-box models, especially when they fail, makes researchers and clinicians question and mistrust their use, thus raising scepticism.

The main objective of this thesis is to make a step forward concerning more effective communication within multidisciplinary teams, which required the joint work of engineering domain techniques, such as signal processing and ML, with tools from social sciences. The analysed patient data comprises scalp recordings provided by the EPILEPSIAE database. This thesis comprises three main contributions.

The first is a sociological study of this research field. Based on the literature, a qualitative study was made to find social barriers concerning the clinical application of seizure prediction algorithms. Two tools were used: while Grounded Theory allows the draw of hypotheses from data, Actor-Network Theory considers that technology shapes social configurations and interests. A social network was obtained, describing this research field ecosystem and proposing research guidelines for clinical acceptance. The most relevant conclusion is the need for model explainability, but not necessarily intrinsically interpretable models. Due to general scepticism, patient safety reasons, and purposely vague legislation on black-box algorithms for high-risk decisions, many authors advocate using only transparent models, limiting their performance and potential. Nevertheless, according to the study conducted in this thesis, researchers may develop robust prediction models, including black-box

systems, to some extent, as long as they can deliver human-comprehensible explanations. This contribution highlights a path, by using model explainability, on how to allow the use of more computationally robust models.

The second contribution is an evolutionary seizure prediction framework whose output is a logistic regression model. The framework identifies the best set of five features (widely explored in the literature) while automatically searching for the preictal period and accounting for patient comfort. It provides patient-specific and patient-general interpretable insights, which might be helpful in better understanding seizure generation processes and explaining the algorithm's decisions. This methodology was quasi-prospectively tested on continuous data, comprising recordings from 93 patients with several types of focal and generalised epilepsies. Performed above chance was achieved for 32% of patients. The results were compared with a seizure surrogate predictor and a control method based on a typical ML pipeline (pre-processing, feature extraction, classification, and post-processing). The obtained findings may evidence the need for patient-specific methodologies.

The third contribution is the evaluation of model explainability by data scientists and epilepsy specialists. Three ML methodologies containing different model transparency levels were developed: a logistic regression, an ensemble of fifteen Support Vector Machines, and an ensemble of three Convolutional Neural Networks. Each methodology was quasi-prospectively evaluated in 40 patients. Patients with high and low performances were chosen to develop explanations. These were presented, during interviews, to data scientists working in healthcare and clinicians from an epilepsy refractory centre. The interviews were analysed and resulted in five lessons leading to better communication of ML models from researchers to clinicians. The most significant finding was that the goal of explainability is not merely to explain the system's decisions but to improve the system itself by questioning the assumptions it is based on. It is hard to understand and interpret brain dynamics even using simple models with state-of-the-art features. In this study, it was possible to conclude that designing several models that explicitly deal with changes in signal dynamics helps develop a complete problem formulation and improve explainability.

Since these contributions used data from pre-surgical monitoring conditions, the obtained findings should be interpreted as proofs of concepts. These methodologies must be replicated in studies using ultra-long-term recordings which concern real-life

conditions.

Keywords: Epilepsy, Seizure Prediction, Electroencephalogram, Interpretability, Explainability, Machine Learning

Resumo

A previsão de crises epilépticas é um problema de análise e classificação de janelas consecutivas ao longo do tempo. Os métodos existentes, que usam o sinal electroencefalográfico, contêm maioritariamente modelos de Machine Learning (ML). Como estes lidam com dados multidimensionais, tornam-se caixas negras, não oferecendo confiança aos clínicos em decisões de alto risco. Apesar de décadas de investigação, poucos dispositivos chegaram a um ensaio clínico ou foram comercializados onde, abordagens mais recentes (como redes neuronais) acabam por não ser totalmente exploradas. A ausência de explicações para modelos considerados caixas negras, especialmente quando estes falham, aumenta o cepticismo dos clínicos.

O principal objectivo desta tese é desenvolver soluções que permitam uma melhor comunicação para equipas multidisciplinares que trabalhem em previsão de crises epilépticas, o que requereu o trabalho conjunto de domínios da engenharia, como processamento de sinal e ML, com ferramentas das ciências sociais. Usaram-se os dados electroencefalográficos de escalpe da base de dados EPILEPSIAE. Esta tese divide-se em três contribuições principais.

A primeira contribuição é um estudo sociológico onde, com base na literatura, se tentou encontrar as barreiras sociais existentes na área de previsão de crises. Usaram-se duas ferramentas: Teoria Fundamentada, que permitiu desenvolver hipóteses a partir dos dados recolhidos, e Teoria Actor-Rede, que considerou o papel activo da tecnologia ao moldar configurações sociais. Obteve-se uma rede social que descreve o ecossistema desta área de investigação, propondo-se directrizes para acelerar o processo de aceitação clínica de modelos matemáticos. A conclusão mais relevante é: apesar de ser necessário explicar os modelos, estes não precisam de ser intrinsecamente interpretáveis. É possível desenvolver modelos com um certo grau de complexidade, desde que consigam explicar as suas decisões de forma humanamente

compreensível. Devido ao cepticismo e motivos baseados numa legislação ainda vaga, muitos investigadores advogam o uso exclusivo de modelos interpretáveis, o que pode limitar a performance e o potencial desta tecnologia. Esta contribuição mostra um caminho possível, através da explicabilidade, para permitir a aplicação clínica de modelos computacionalmente mais complexos.

A segunda contribuição é uma metodologia evolucionária de previsão de crises que devolve um modelo de regressão logística interpretável. Esta metodologia identifica o melhor conjunto de cinco preditores enquanto procura o período pré-ictal e considera o conforto do doente. Ao fornecer conhecimento especializado para cada doente e generalizado para todos os doentes, pode contribuir para um melhor entendimento acerca do processo de geração de crises e das decisões de cada modelo. Testou-se esta metodologia quase-prospectivamente em dados contínuos, pertencentes a 93 doentes que sofrem de vários tipos de epilepsias focais e generalizadas. Esta metodologia teve significância estatística para 32% dos doentes através de uma análise *surrogate*, e foi comparada com um método de controlo baseado na literatura (pré-processamento, extracção de características, classificação, pós-processamento). Os resultados evidenciam a necessidade de desenvolver diferentes metodologias para diferentes modelos.

A terceira contribuição consistiu na avaliação de explicabilidade de modelos por parte de especialistas em ML e epilepsia. Desenvolveram-se três metodologias de ML: uma regressão logística, um sistema de voto de 15 *Support Vector Machines*, e um sistema de voto de três Redes de Convolução Neuronais. Testaram-se as metodologias quase-prospectivamente em 40 doentes. Seleccionaram-se doentes com alta e baixa performance para se desenvolverem explicações acerca dos respectivos modelos. Estas explicações foram apresentadas, em entrevista, a cientistas de dados que trabalham em problemas clínicos e a clínicos de um centro de epilepsia refractária. Da análise das entrevistas, resultaram cinco lições para melhorar a comunicação de modelos de ML. A descoberta mais significativa foi que o objectivo da explicabilidade não é só explicar as decisões do sistema, mas melhorar a metodologia em si. Foi ainda importante reforçar que, apesar de se terem utilizado preditores clinicamente intuitivos e conhecidos do estado da arte juntamente com modelos simples, foi difícil interpretar os resultados e obter conhecimento acerca das dinâmicas cerebrais. Pode-se atingir um aumento da explicabilidade ao desenvolverem-se, em

paralelo, sistemas que lidem explicitamente com mudanças nas dinâmicas cerebrais que ajudem a obter uma definição mais completa do problema.

Como estas contribuições usaram dados de monitorização pré-cirúrgica, os resultados obtidos deverão ser interpretados como uma prova de conceito. Estas metodologias deverão ser replicadas com dados de longa duração da vida real.

Keywords: Epilepsia, Previsão de Crises, Electroencefalograma, Interpretabilidade, Explicabilidade, Machine Learning

Contents

List of Figures

List of Tables

List of Acronyms

AEDs Anti-Epileptic Drugs 1, 13, 18, 19, 23, 39, 45

ANT Anterior Nucleus of the Thalamus 21, 23

AUC Area Under the Curve 60, 224

CD Concept Drift xxii, 2, 27, 33, 36, 37, 61, 65, 251

CNN Convolutional Neural Networks 55, 57, 60, 81, 108–112, 122, 241, 250

DBS Deep Brain Stimulation 21, 23

DL Deep Learning xxii, 48, 49, 55, 57, 60, 63–65, 68, 80, 82, 101, 107

DM Decision Maker 92

DRE Drug-Resistant Epilepsy xxi, 1, 18, 19, 22, 30, 44, 65, 69, 73, 75, 108, 173, 174, 185, 192, 197

EA Evolutionary Algorithms 4, 9, 83–85, 101, 221

EEG Electroencephalogram xxi, xxix, 1, 2, 4, 9, 11, 13–19, 23–29, 31, 33, 34, 36–39, 44, 45, 47–49, 51, 53, 60, 62–64, 67, 73, 76, 79, 81, 84–87, 102, 107–109, 111, 115, 116, 119, 121–123, 125, 126, 161–164, 174–176, 185, 193, 256

FBTC Focal to Bilateral Tonic-Clonic 11, 13

FDA Food and Drug Administration 23, 76

FFT Fast Fourier Transform 60

FOA Focal Onset Aware 11, 13, 94, 95

FOIA Focal Onset Impaired Aawareness 11, 13, 94, 95

FPR/h False Prediction Rate per hour xxix, 28–32, 45, 60, 61, 79, 92–95, 99–103, 109, 110, 114, 193, 197

GDPR General Data Protection Regulation 2, 39, 74, 82, 101, 122, 167, 182, 185

GT Grounded Theory 9, 67, 70, 108, 113, 123

iEEG Invasive Electroencephalogram 14, 17, 18, 44, 49, 174, 175, 185, 187, 223

ILAE International League Against Epilepsy 10, 12, 13, 18, 19

kNN k-Nearest Neighbours 40, 57, 77

LIME Local Interpretable Model-Agnostic Explanations 42, 112

LSTM Long Short-Term Memory 55, 57, 60, 84

ML Machine Learning vii, viii, xi, xii, xxiv, 2–4, 25, 32, 38–40, 43–45, 47, 48, 51, 64, 69, 70, 77, 80, 82–84, 92, 104, 107, 108, 110, 117, 118, 121–123, 159

MOEA Multiobjective Evolutionary Algorithm xxiii, xxv, xxvi, xxix, 85–95, 99–104, 161, 209, 211, 215, 219, 221, 223–225, 229–231, 233–235, 256

MRI Magnetic Resonance Imaging 19, 20, 126

NREM Non-Rapid Eye Movement 38

NSGA-II Non-dominated Sorting Genetic Algorithm II 87, 88

PSD Power Spectral Density 162

RBF Radial Basis Function 59

REM Rapid Eye Movement xxii, 38

RFE Recursive Feature Elimination 224

RNN Recurrent Neural Network 60, 84

RNS Responsive Neurostimulation System 2, 21, 23, 185, 187, 191, 194, 197

ROC Receiver Operating Characteristic 29, 224

SEF Spectral Edge Frequency 97

SHAP SHapley Additive exPlanations 64, 111–113

SOP Seizure Occurrence Period xxi, xxii, 28–30, 32, 33, 35, 36, 45, 51, 54, 76, 187, 193

SPH Seizure Prediction Horizon xxi, xxix, 28–30, 36, 45, 51, 52, 54, 76, 187, 193, 211

SVM Support Vector Machines xxiv, 57, 59, 64, 108, 110, 111, 114, 117, 118, 122, 224, 241, 249, 250

TiW Time in Warning 61, 121, 122

TLE Temporal Lobe Epilepsy 13, 17, 21, 44

VNS Vagus Nerve Stimulation 21, 23

Chapter 1

Introduction

The last four decades have seen significant advances in seizure prediction using the Electroencephalogram (EEG) signal. Prediction models must deal with the inherent complexity of epilepsy and its seizures, where their decisions might significantly impact the patients' lives. Thus, researchers must improve communication strategies to explain the models' decisions, particularly to patients and clinicians. This chapter addresses the need for compelling explanations concerning the mathematical models' decisions. On that basis, it traces the main goals of this thesis.

1.1 Motivation

Epilepsy is one of the most common neurological diseases. While prevailing in approximately 1% of the world population, it affects people of all ages and conditions [Ihle et al., 2012]. The occurrence of apparently unpredictable seizures is the most severe aspect, impacting the lives of patients and caregivers. While delivering Anti-Epileptic Drugs (AEDs) has a success rate of 70% in achieving seizure control, Drug-Resistant Epilepsy (DRE) patients require other strategies, such as seizure prediction, to improve their lives [Jette and Engel, 2016, Cloppenborg et al., 2016, Klatt et al., 2012]. Without proper seizure control, the life of a patient with DRE is significantly affected due to not only discrimination and stigma but also to economic reasons regarding health care needs, premature and/or sudden death, loss of productivity, depression, and anxiety [Fiest et al., 2017, Jones and Thomas, 2017, Devinsky et al., 2016, Laxer et al., 2014].

Seizure prediction is a promising research field as it offers several solutions. The design of a warning device that timely anticipates a seizure, by analysing the EEG in real-time, may allow the patient to take action for seizure suppression (by administering rescue medication) and/or to minimise its effects (by taking preventive measures against accidents). It is also possible to integrate a prediction model in a closed-loop system that automatically performs neuromodulation to suppress

seizures [Kuhlmann et al., 2018b, Freestone et al., 2017, Mormann et al., 2007]. Researchers have proven that these solutions are possible in real life, with the Neurovista's Seizure Advisory System Feasibility Study (NCT01043406) [Cook et al., 2013] and the Responsive Neurostimulation System (RNS) Long-Term Treatment (NCT00572195) [Sun and Morrell, 2014] clinical trials, for instance. However, not all patients from the Neurovista clinical trial reached the final phase, and, in those who did, not all achieved good performance. Although the use of the RNS® System significantly reduced seizure occurrence, complete seizure suppression was not possible. Also, these systems are highly invasive, demanding high monetary costs and multidisciplinary teams, which may not be accessible in all countries [Engel, 2016].

1.1.1 The importance of explaining decisions

Nowadays, seizure prediction encompasses mainly the development of Machine Learning (ML) models using multidimensional time-series data. As a result, seizure prediction models are often considered highly complex black-box models. The absence of explanations for black-box models' decisions, especially when they fail, leads researchers and clinicians to question and mistrust their use, thus raising scepticism [Freestone et al., 2017]. If one tries to explain a model's decision, particularly when it failed (why it missed a seizure or raised a false alarm), it might convince a clinician that it brings added value [Molnar, 2019, Lage et al., 2019, Doshi-Velez and Kim, 2017]. Additionally, focusing on explaining these models' decisions can also provide ways of understanding how to improve methodologies.

Explaining models' decisions is vital for any clinical problem, as it can be understood when analysing the 2018 General Data Protection Regulation (GDPR) [Goodman and Flaxman, 2017, Doshi-Velez and Kim, 2017] and the 2021 European Union Medical Device Regulation (EU MDR) [Beckers et al., 2021, Majety et al., 2021] for European citizens and the European economic space. GDPR 's article 22 presents the first steps toward legislation on algorithm explainability for high-risk decisions based on personal data. It provides patients with the right to have an explanation for any algorithm decision. Also, it gives them the right to question those decisions. The 2021 EU MDR also promotes the delivery of model explanations.

1.1.2 Seizure prediction limitations

Seizure prediction is a research field whose success is severely affected by the heterogeneity of seizures and epilepsies, a significant data imbalance resulting from the rare occurrence of seizures, and Concept Drift (CD) [Kuhlmann et al., 2018b, Baud et al., 2018, Karoly et al., 2017, Freestone et al., 2017, Gadhoumi et al., 2016a]. It is possible to outline several reasons that may explain the lack of adequate performance for all patients, where the failure of applying current EEG-based approaches to real-life may be related to the quality of databases. Although some contain continuous and

long-term recordings, these require context information such as epilepsy type, focus location and lateralisation, vigilance state, medication, and others to deepen current knowledge on the seizure generation process [Kuhlmann et al., 2018b].

Current state-of-the-art methods focus on supervised learning techniques that distinguish chunks of chronological information as either seizure-free (interictal) state or pre-seizure (preictal) state [Kuhlmann et al., 2018b, Mormann et al., 2007]. Correct labelling of the preictal interval is critical to achieving good performance. This interval is one of the fundamental aspects of seizure prediction, with literature providing evidence of its existence. A consensus regarding an optimal preictal period was not achieved, with authors reporting different optimal values [Kuhlmann et al., 2018b, Bandarabadi et al., 2015a]. Studies attempting to determine the preictal interval verify that this transitional stage varies from subject to subject and from seizure to seizure within patients, which explains the tendency to develop patient-tailored prediction algorithms [Kuhlmann et al., 2018b, Freestone et al., 2017].

1.2 Goals and contributions

The main objective of this thesis is to provide novel solutions for ML seizure prediction models in such a way that it contributes to its interpretation and thus allow more effective communication within multidisciplinary teams, namely among data scientists and clinicians. This thesis required the joint work of engineering domain techniques, such as signal processing and ML, with tools from social sciences to handle the complexity of this communication. The used dataset comprises recordings provided by the European Epilepsy Database [Klatt et al., 2012]. Since all data was recorded during pre-surgical monitoring, the findings herein should be interpreted as proof of concept. This research can bring added value to future studies proposing ML models based on data acquired in real-life conditions.

This investigation can be subdivided into three main contributions, described in the following subsections.

1.2.1 Studying the seizure prediction research ecosystem

The first part of this thesis concerns a sociological study of this research field. Despite being useful for clinicians and patients to understand this ecosystem, this study is directed to researchers so that they can develop prediction approaches with a higher chance of clinical acceptance. Thus, this study proposed some guidelines to improve clinical acceptance of seizure prediction algorithms. Although some researchers implicitly use these guidelines, there is added value in explicitly addressing and discussing them. A research field's social network was also created, along with an interactive presentation to illustrate this ecosystem better.

1.2.2 Interpretable evolutionary algorithms

The second part of this thesis presents a patient-specific methodology for scalp EEG seizure prediction, tested quasi-prospectively, which returns a classification model with high transparency and requires low-computational power. The model is a logistic regression that uses five features widely explored in the literature. This methodology uses Evolutionary Algorithms (EA), which are based on a population of individuals (points in the search space) and are inspired by natural evolution. They are helpful for direct search, optimisation, and ML problems. Briefly, the individuals, defined by a set of features that best performed in seizure prediction, survived and proliferated. Then, further analysis of the obtained solutions was performed to search for patient-specific and patient-general behaviours. It was also possible to address patient comfort during data acquisition by assessing the selected electrodes (number and spatial distribution in the scalp).

1.2.3 Explaining Machine Learning prediction models

The third part of this thesis concerns the explainability evaluation of different ML methodologies with different levels of transparency. For each methodology, prediction performance was quasi-prospectively obtained. Then, patients with good and poor performances were chosen to explain the models' prediction decisions. The developed explanations were presented (in interviews) to data scientists working on clinical problems and clinicians working in an epilepsy refractory centre. These interviews were analysed and discussed, resulting in five lessons that lead to better communication of ML models to clinicians.

1.3 Thesis Outline/Structure

The remainder of this thesis proposal is structured as follows.

Chapter 2 provides background information related to epilepsy, the EEG signal, an introduction to the field of seizure prediction and ML explainability.

Chapter 3 presents a literature overview on EEG seizure prediction and EEG-based models' explainability, while presenting major current limitations.

Chapter 4 analyses the seizure prediction social ecosystem, along with the obtained social network and proposed guidelines.

Chapter 5 presents the development of interpretable EAs for EEG seizure prediction.

Chapter 6 describes the explainability study for different ML approaches, from intrinsically interpretable models to black-box ones.

Chapter 7 concludes this thesis by outlining its main findings and highlighting their added value. It also presents future directions.

1.4 Scientific contributions

During this thesis, several contributions to epilepsy seizure prediction research were made. These comprise publications as main-author and co-author in international journals, participating in international and national conferences, as masters' degree theses co-advisor, and participation in science communication sessions to the general public. These are enumerated here.

Educational and scientific contributions in other research fields were also made. These include giving classes and workshops to students, invited speaker and attendee in summer schools, science communication activities to the general public, masters' theses advising, and publications in international journals (see Appendix A for full detail on these contributions).

1.4.1 Articles in international journals

J1 **Pinto, M. F.**, Leal, A., Lopes, F., Dourado, A., Martins, P., and Teixeira, C. A.. "A personalized and evolutionary algorithm for interpretable EEG epilepsy seizure prediction", *Scientific Reports* 11, 3415, DOI: 10.1038/s41598-021-82828-7 (2021).

J2 **Pinto, M. F.**, Coelho, T., Leal, A., Lopes, F., Dourado, A., Martins, P., Teixeira, C. A.. "Interpretable EEG seizure prediction using a multiobjective evolutionary algorithm", *Scientific Reports*, 12, 4420, DOI: 10.1038/s41598-022-08322-w (2022).

J3 **Pinto, M. F.**, Leal, A., Lopes, F., Pais, J., Dourado, A., Sales, F., Martins, P., Teixeira, C. A.. "On the clinical acceptance of black-box systems for EEG seizure prediction", *Epilepsia Open*, DOI: 10.1002/epi4.12597 (2022).

J4 Leal, A., **Pinto, M. F.**, Lopes, F., Bianchi, A. M., Henriques, J., Ruano, M. G., Carvalho, P., Dourado, A., and Teixeira, C. A.. "Heart rate variability analysis for the identification of the preictal interval in patients with drug-resistant epilepsy", *Scientific Reports* 11, 5987, DOI: 10.1038/s41598-021-85350-y (2021).

J5 Lopes, F., Leal, A., Medeiros, J., **Pinto, M. F.**, Dourado, A., Dümplemann, M., Teixeira, C. A.. "Automatic Electroencephalogram Artifact Removal using Deep Convolutional Neural Networks", *IEEE Access* 9 149955-149970, DOI: 10.1109/ACCESS.2021.3125728 (2021).

J6 Lopes, F., Leal, A., Medeiros, J., **Pinto, M. F.**, Dourado, A., Dümplemann, M., Teixeira, C. A.. "Ensemble Deep Neural Network for Automatic Classification of EEG Independent Components", *IEEE Transactions on Neural*

Systems and Rehabilitation Engineering 30 559-568, DOI: 10.1109/TNSRE.2022.3154891 (2022).

J7 Lopes, F., Leal, A., Medeiros, J., **Pinto, M. F.**, Dourado, A., Dümplemann, M., Teixeira, C. A.. "EPIC: Annotated epileptic EEG independent components for artifact reduction", *Scientific Data* 9 512, DOI: https://doi.org/10.1038/s41597-022-01524-x (2022).

1.4.2 Articles under preparation to be submitted to international journals

J8 **Pinto, M. F.**, Batista, J., Leal, A., Lopes, F., Oliveira, A., Dourado, A., Sales, F., Martins, P., Teixeira, C. A.. "Explaining Machine Learning models for EEG seizure prediction", *Manuscript under preparation to be submitted to a scientific journal* (2022).

J9 Leal, A., Curty, J., Lopes, F., **Pinto, M. F.**, Oliveira, A., Sales, F., Ruano, M. G., Carvalho, P., Dourado, A., Henriques, J., and Teixeira, C. A.. "Unsupervised EEG Preictal Interval Identification in Patients with Drug-resistant Epilepsy." *Pre-print.* DOI: https://www.researchsquare.com/article/rs-1905838/v1 (2022).

J10 Leal, A., Martinho, B., Lopes, F., **Pinto, M. F.**, Sales, F., Bianchi, A.M., Ruano, M. G., Dourado, A., Henriques, J., and Teixeira, C. A.. "Preictal Interval Labelling with Supervised Learning May Improve Seizure Prediction Models", *Manuscript under preparation to be submitted to a scientific journal* (2022).

1.4.3 Participation in conferences

C1 **Pinto, M. F.**, Leal, A., Lopes, F., Dourado, A., Martins, P., and Teixeira, C. A.. "Can we explain how Machine Learning Models predict seizures? Towards an appropriate explainability of EEG seizure prediction models" in *International Conference for Technology and Analysis of Seizures, 2022 (ICTALS 2022).*

C2 **Pinto, M. F.**, Leal, A., Lopes, F., Pais, J., Dourado, A., Sales, F., Martins, P., Teixeira, C. A.. "Searching for the Epilepsy Seizure Prediction Ecosystem" in 33º National Encounter of Epileptology (ENE) – Virtual Congress of the Portuguese League Against Epilepsy Virtual (LPCE).

1.4.4 Master's degree theses co-advisoring

M1 Coelho, T.. "EEG Epilepsy Seizure Prediction: A Multi-Objective Evolutionary Approach", *Master Thesis dissertation, Universidade de Coimbra* (2020).

M2 Tavares, M.. "EEG Epilepsy Seizure Prediction: the Post-Processing Stage as a Chronology", *Master Thesis dissertation, Universidade de Coimbra* (2021).

M3 Oliveira, A. C.. "Sleep-Awake cycle evaluation from long-term EEG data: assessing the impact in epilepsy seizure prediction", *Master Thesis dissertation, Universidade de Coimbra* (2021).

M4 Pontes, E. D.. "Concept-Drifts Estimation for EEG Epilepsy Seizure Prediction", *Master Thesis dissertation, Universidade de Coimbra* (to be finished in 2022).

M5 Batista, J.. "On the development of EEG seizure prediction methodologies aimed at clinically acceptance", *Master Thesis dissertation, Universidade de Coimbra* (to be finished in 2022).

1.4.5 Science communication to the general public

G1 "An Evolutionary Framework for Rare-Event Prediction Problems with Machine Learning", in Data Science Portugal (DSPT) meetup #68, Coimbra (2019).

G2 "Can your artificial intelligence algorithm predict epileptic seizures and explain it to a doctor?", 3 Minute-Thesis Competition (3MT) from the University of Coimbra (2020).

G3 "Anticipating and disarming epileptic seizures: utopia or future?", in International Science Festival (FICA), Oeiras, Lisbon (2021).

1.4.6 Awards and distinctions

A1 The abstract "Can we explain how Machine Learning Models predict seizures? Towards an appropriate explainability of EEG seizure prediction models" was considered one of the top-ranked abstracts in the ICTALS 2022 conference.

A2 The paper "A personalized and evolutionary algorithm for interpretable EEG epilepsy seizure prediction", from *Scientific Reports*, is in the *2021 Top 100 in Neuroscience*. More specifically, it was the 32$^{\text{nd}}$ most downloaded paper from all the neuroscience articles published in 2021 in *Scientific Reports*.

A3 The pitch "Can your artificial intelligence algorithm predict epileptic seizures and explain it to a doctor?" was the winner of the 3MT competition in the University of Coimbra (2020).

Chapter 2

Background concepts

This chapter introduces the main background concepts. Section 2.1 presents a brief notion of epilepsy and seizures. Section 2.2 describes the Electroencephalogram (EEG) signal along with its materialisation in neurophysiology and epilepsy. Section 2.3 details the current treatment and therapeutics options. Section 2.4 provides a brief introduction to the theoretical concepts behind seizure prediction. Section 2.5 presents some concept drifts important for seizure prediction and section 2.6 provides a summary of explainability and interpretability concepts. Finally, section 2.7 provides a summary and discussion of the background key concepts.

This thesis also requires understanding Evolutionary Algorithms (EA) and qualitative research tools, such as Grounded Theory (GT) and Actor-Network Theory. As it may be more intuitive, these are briefly introduced during their use in chapters 4 and 5, respectively.

2.1 Epilepsy and seizure concepts

Epilepsy has a worldwide prevalence of 1% and is one of the most common neurological diseases with profound physical, psychological, and social consequences. Recurrent and typically brief episodes, known as seizures, characterise this disease [Van Mierlo et al., 2014, Iasemidis, 2003].

Excessive electrical discharges in cells from one or more brain parts are responsible for causing seizures, ranging from the briefest lapses of attention or muscle jerks to severe and prolonged convulsions. These may differ in frequency and severity, ranging from less than one per year to several per day. The clinical manifestations of seizures may present many forms which depend on the affected areas, including auras, tonic-clonic movements, impairment, or loss of consciousness [Van Mierlo et al., 2014, Bautista and Glen, 2009, Fisher et al., 2005].

2.1.1 Definition of epilepsy and seizures

According to the International League Against Epilepsy (ILAE) in 2005 [Fisher et al., 2005], epilepsy and an epileptic seizure could be defined as presented in Boxes 1 and 2, respectively.

Box 1 - Definition of epilepsy according to the 2005 ILAE Task Force [Fisher et al., 2005].

"Epilepsy is a disorder of the brain characterised by an enduring predisposition to generate epileptic seizures, and by the neurobiologic, cognitive, psychological, and social consequences of this condition. The definition of epilepsy requires the occurrence of at least one epileptic seizure."

Box 2 - Definition of seizure according to the 2005 ILAE Task Force [Fisher et al., 2005].

"An epileptic seizure is a transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain."

More recently, in 2014, the ILAE proposed an operational clinical definition of epilepsy [Fisher et al., 2014] to be applied in diagnostics, which can be seen in Box 3. This new practical definition aims to raise awareness among clinicians to the risk of recurrence after a single unprovoked seizure, which may enable an earlier treatment start.

Box 3 - Operational clinical definition of epilepsy according to the 2014 ILAE Task Force [Fisher et al., 2014].

According to this update, epilepsy is a disease of the brain defined by any of the following conditions:

- *at least two unprovoked (or reflex) seizures occurring 24h apart;*
- *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 ten years;*
- *diagnosis of an epilepsy syndrome.*

2.1.2 Classification of seizures and epilepsies

In 2017, the ILAE also updated the framework for classification of the epilepsies [Scheffer et al., 2017] and the operational classification seizures types [Fisher et al., 2017] (see Figure 2.1). Three stages comprise the framework for the classification of epilepsies: (i) seizure type, (ii) epilepsy type, and (iii) epilepsy syndrome. The diagnostic process may include the assessment of medical history and physical examination. Classification according to seizure type may be the maximum level

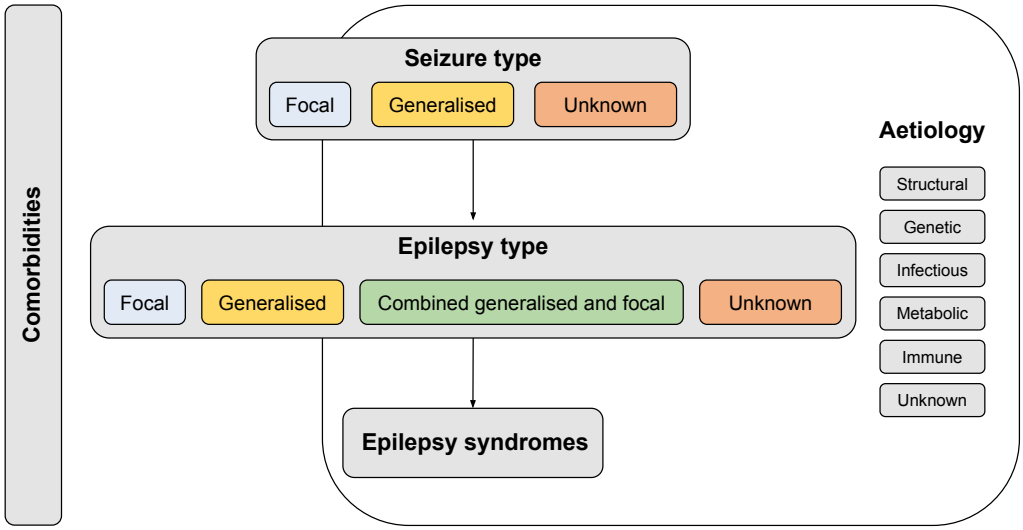


Figure 2.1: ILAE 2017 classification of epilepsies. Adapted from [Scheffer et al., 2017].

for diagnosis as there may be no access to video-EEG and neuroimaging examination [Elger and Hoppe, 2018].

This new classification incorporates aetiology along each stage, emphasising the need to consider it at each step as it often carries significant treatment implications. Despite being generally unknown, aetiology can be divided into several categories: structural, genetic, infectious, metabolic, and immune. Comorbidities such as learning, psychological, and behavioural problems are also important factors to consider for patient management [Devinsky et al., 2018, Scheffer et al., 2017, Fisher et al., 2014]. This thesis did not explore aetiology influence and comorbidities.

2.1.2.1 Seizure types

An epilepsy diagnosis starts by discovering the type of seizures a patient suffers from, which concerns how and where they begin in the brain. A seizure can be classified into focal, generalised or unknown [Devinsky et al., 2018, Scheffer et al., 2017], as seen in Figure 2.2.

Limiting a clinical and EEG onset moment to one cerebral hemisphere is possible in focal seizures. In contrast, with generalised seizures, both hemispheres are involved. A generalised onset is characterised by the *"engagement of bilateral networks"* which may not be necessarily symmetric. Possible subcategorisation for both focal and generalised seizures concerns the prominent symptom during a seizure into motor or non-motor. It is still possible to classify a seizure in terms of awareness of self and the environment. Regarding awareness, it is classified as Focal Onset Aware (FOA) or as Focal Onset Impaired Aawareness (FOIA). Also, there is the particular case of Focal to Bilateral Tonic-Clonic (FBTC). These often lead to tonic (body stiffness) and clonic (jerking movements) symptoms. Although their onset is limited to one hemisphere, they quickly propagate to another brain region [Devinsky et al.,

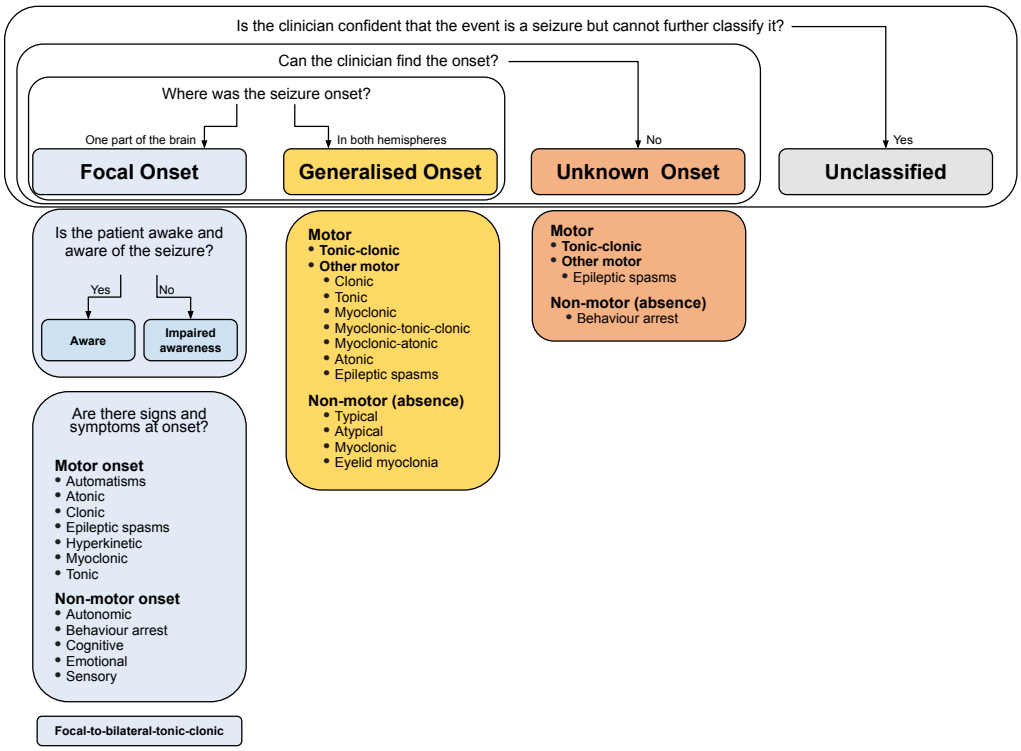


Figure 2.2: ILAE 2017 classification of seizure types. Clinicians may use a basic (in bold) or an expanded seizure classification, which depends on their expertise in diagnosing and treating epilepsy. Adapted from [Devinsky et al., 2018, Fisher et al., 2017].

2018, Scheffer et al., 2017].

It is essential to mention generalised tonic-clonic seizures, a type of seizure involving the entire body, also known as *grand mal* seizures. The terms seizure, convulsion, or epilepsy are often associated with generalised tonic-clonic seizures as it is the most common seizure type [Kodankandath et al., 2020]. When defining a precise onset localisation is impossible, the seizure onset is classified as unknown.

Even though the ILAE does not have classification guidelines for more aspects, it is common and useful to analyse and label other related factors to provide context information [Fisher et al., 2017]. This thesis often stratifies patients by their vigilance state at the time of the seizure (awake, REM, nonREM stages I, II, III, or IV), the seizure onset place in terms of brain lobes (frontal, temporal, central, occipital, and parietal), and in terms of hemispheres (left, right, and bilateral).

2.1.2.2 Epilepsy type

The next step identifies the type of epilepsy based on the patient's seizures. It assumes the patient has epilepsy based on the 2014 definition and includes an unknown category when the clinician does not have enough information. Furthermore, there is a large complexity associated with an epilepsy type as each category includes multiple types of seizures. As one can see with Figure 2.1, there are four types of

epilepsies [Devinsky et al., 2018, Scheffer et al., 2017]:

- Focal: include unifocal and multifocal disorders and seizures that involve one hemisphere. A patient with focal epilepsy can have different types of seizures. The most frequent are FOA, FOIA, focal motor, focal non-motor, and FBTC ones. The interictal EEG often contains focal epileptiform discharges.
- Generalised: seizures start on both hemispheres. It is characterised by a spike-wave activity on the EEG, and the most frequent seizure types are absence, myoclonic, atonic, tonic, and tonic-clonic ones.
- Combined Generalised and Focal: characterised by having generalised and focal seizures where the interictal EEG shows generalised spike-wave and focal epileptiform discharges.

Concerning focal epilepsies, Temporal Lobe Epilepsy (TLE) is the most common type and is characterised by focal seizures which arise in the temporal lobe. TLE can be further divided into mesial, neocortical, or lateral. This fact is essential since the majority of patients diagnosed with mesial TLE become resistant to Anti-Epileptic Drugs (AEDs) and, therefore, are referred to resective surgery [Anyanwu and Motamedi, 2018, Varsavsky et al., 2011]. Since this work concerns the exploration of the EPILEPSIAE database [Klatt et al., 2012], solely constituted by patients in presurgical monitoring, the majority of the studied patients has TLE.

2.1.2.3 Epilepsy syndrome

An epilepsy syndrome consists of a cluster of characteristics that incorporates seizure types, EEG, and imaging features which tend to occur together. Furthermore, age at onset and remission, seizure triggers, diurnal variation, and others can also be considered. It is worth noting that there is no formal classification of syndromes by the ILAE [Scheffer et al., 2017].

Identifying an epilepsy syndrome may help determine the underlying causes of seizures and help decide which medication a patient should take, thus serving a different purpose such as guiding epilepsy treatment. This identification is crucial as some syndromes demonstrate seizure aggravation with the incorrect anti-seizure medication, which can be avoided with the appropriate diagnosis [Gayatri and Livingston, 2006].

2.1.3 Seizure clusters

Despite not being listed in the ILAE classification, a seizure cluster is a relevant phenomenon known for acute repetitive seizures within short interictal intervals (hours or minutes). In the literature, it is frequent to report seizure clusters as consecutive seizures within a determined period, which varies from study to study.

This is particularly important for patients in presurgical monitoring and seizure prediction [Jafarpour et al., 2019, Mormann et al., 2007], as it will be seen in the Treatment and Therapeutics (2.3) and Seizure Prediction (2.4) sections.

2.2 EEG

The EEG captures the brain electrical activity in a time series. This activity is the result of the potential voltage of the summed excitatory and inhibitory potentials produced by brain cells and their geometrical disposition [Mporas et al., 2015, Alotaiby et al., 2014, Iasemidis, 2003]. The acquisition of the EEG signal is performed by placing electrodes (electrical sensors) in the scalp (scalp EEG) or inside the skull (Invasive Electroencephalogram (iEEG)). It is the most efficient medical imaging tool to analyse the characteristics of this neurological disorder [Medithe and Nelakuditi, 2016]. The EEG is a nonlinear and nonstationary signal, considered complex and challenging to interpret [Acharya et al., 2013].

The number of electrodes and the correspondent localisation determine the signal spatial resolution, while the sampling frequency determines the time resolution. Although other neuroimaging techniques may have a better spatial resolution, the EEG provides the best temporal resolution of the cortical function. The heterogeneity of clinical manifestations is related to the different seizure-engaged brain areas which suffer such discharges. The EEG has been used for presurgical evaluation, continuous monitoring, and diagnosis [Feyissa et al., 2021].

2.2.1 EEG activity

The EEG captures two types of potentials (see Figure 2.3): oscillations and transients.

Oscillations can be described as rhythmic fluctuations caused by mechanisms within individual neurons or interactions between them. Based on this, the macroscopic neural oscillations are usually characterised in terms of frequency band activity: delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-128 Hz) that can be divided in sub-bands [Wang et al., 2018,Bandarabadi et al., 2015b,Adeli et al., 2007]. These are associated with different human activities. Delta oscillations are associated with deep sleep, while theta rhythms are observed during drowsiness, creative inspiration, and deep meditation. Alpha waves are the most proeminent brain rhythm, usually observed over the occipital lobe. Beta oscillations usually appear not only during mental and cognitive tasks but also in anxious and alert states, especially in the frontal and central brain regions. Gamma oscillations are rare and are usually clouded in the presence of muscle artefacts, particularly with scalp EEG [Sanei and Chambers, 2007].

Transients can be normal or abnormal. Normal ones are related to sleep poten-

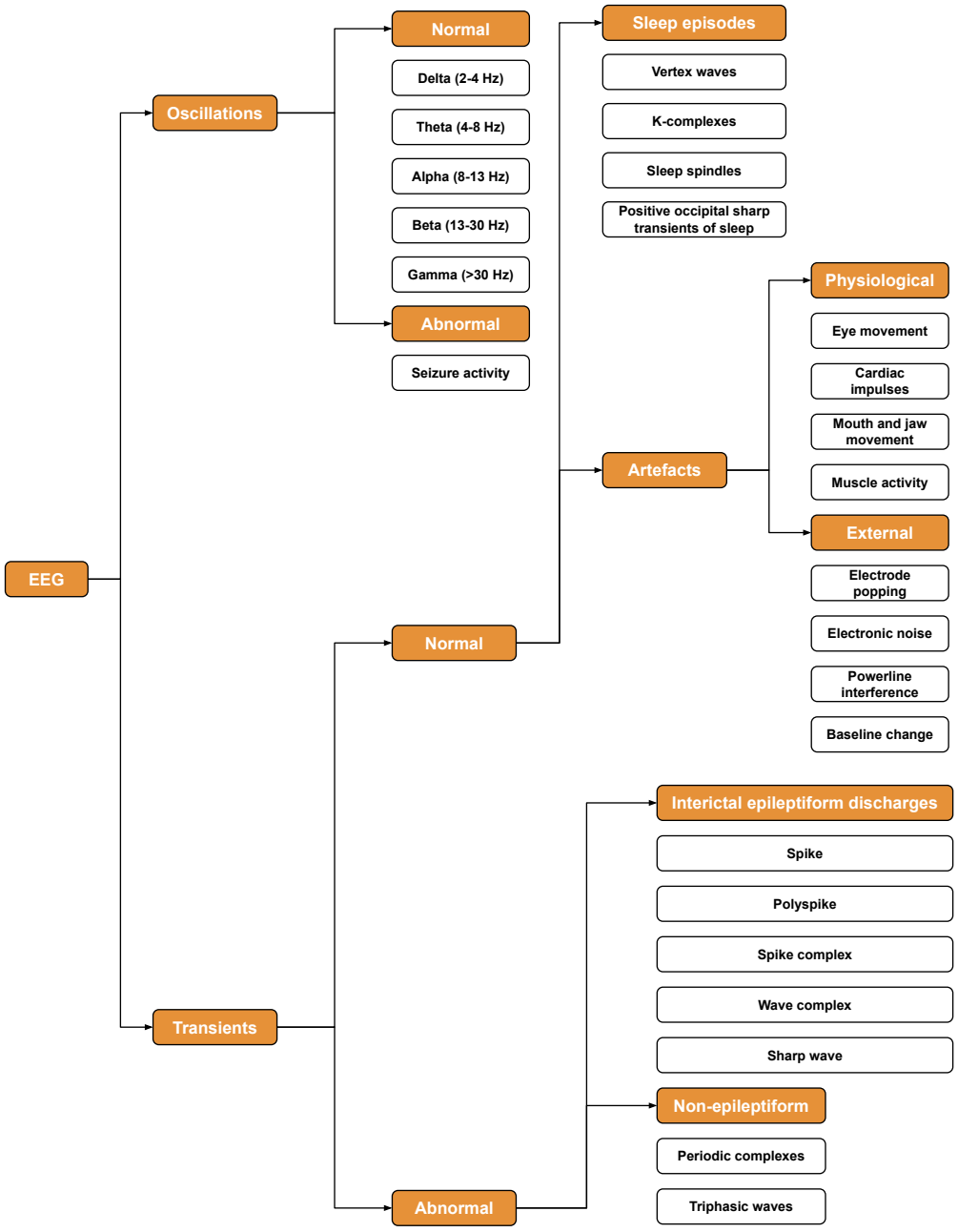


Figure 2.3: EEG activity. Adapted from [Sanei and Chambers, 2007].

tials and biological artefacts such as eye blinking, cardiac, and muscular impulses. As the amplitude of these artefacts is significantly larger compared to regular activity, expert knowledge is required to interpret the EEG signal. Environmental artefacts can also be found. The most common are the 50 or 60 Hz ground frequency, intravenous therapy, and momentary changes in the electrode impedance [Oliveira et al., 2016].

Abnormal transients can be divided into epileptiform and non-epileptiform activity and focal or diffuse. A focal epileptiform activity represents fast, sharp, and synchronous potentials in a significant number of neurons in a discrete brain re-

gion. This activity is sometimes present in interictal periods and may indicate that a given region predisposes to seizure-generation processes. Generalised epileptiform discharges are synchronous and are present in the entire brain, suggesting a generalised epilepsy [Owolabi et al., 2018, Medithe and Nelakuditi, 2016].

2.2.1.1 Scalp EEG

Scalp EEG, or simply EEG, is the most common brain signal as it is minimally invasive. Each electrode records the potential difference from its region to a reference electrode and represents a channel. Usually, electrodes are placed according to the International 10-20 system on the scalp using a conductive gel to decrease impedance (see Figure 2.4.a) [Wennberg, 2011, Varsavsky et al., 2011, Jurcak et al., 2007].

There are three possible montages: i) bipolar (Figure 2.4.b), where the reference electrode is adjacent to each electrode, ii) the referential (Figure 2.4.c), where one electrode is selected as a reference for all, and the average, that has as reference the average potential of all electrodes. Each montage may allow a different analysis, as the reference selection influences the results due to artefacts in different channels [Anastasiadou et al., 2019, Varsavsky et al., 2011, Nunez and Srinivasan, 2006].

Detecting the electrical activity from specific structures within the cortex, such as mesial temporal regions, interhemispheric frontal lobe, lobe structures, and thalamus, is very difficult. In addition, low amplitude fast oscillations in the beta and gamma bands are often contaminated by extracranial (mainly muscle) arte-

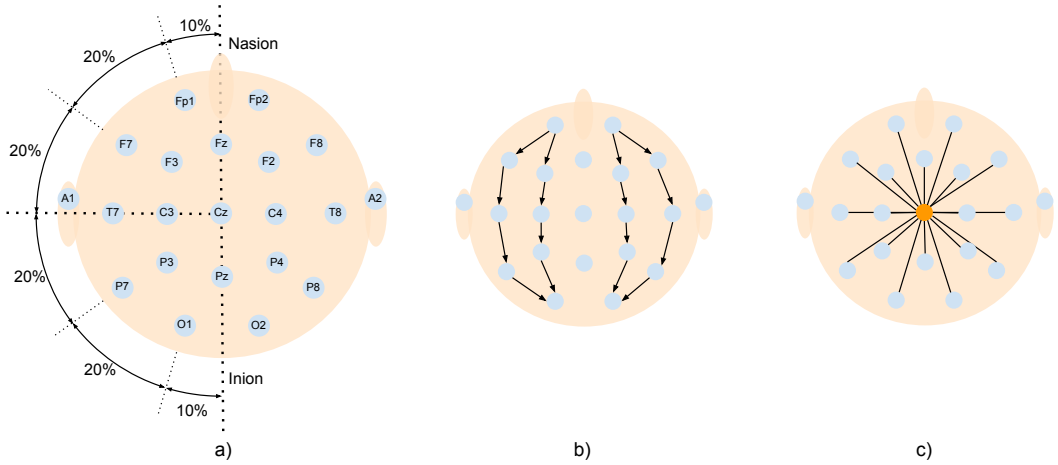


Figure 2.4: Through pre-amplifiers, the EEG electrodes transform ionic current into an electrical one, which is recorded over time. EEG 10-20 electrode placement (a), where 19 electrodes plus one for grounding are placed at 10%, 20%, 20%, 20%, 20% and 10% from inion to nasion along its length. Recently, the international placement systems have renamed four electrodes: T3 to T7, T4 to T8, T5 to P7 and T6 to P8. Letters correspond to the brain lobe where the electrode is placed (F: frontal, T: temporal, P: parietal or posterior temporal, O: occipital, and A: auricular). The numbers assigned to electrodes increase with the distance to the midline. The left and right sides are indicated by odd and even numbers, respectively. The electrodes in the midline are identified with the suffix "z". Regarding montages, it is possible to see a bipolar (b) and a referential one (c). Adapted from [Varsavsky et al., 2011].

facts, as these tend to be less synchronous over large cortex regions [Wennberg, 2011, Varsavsky et al., 2011].

Concerning ambulatory conditions [Biondi et al., 2022], there are already several acquisition systems using scalp EEG that try to provide higher patient comfort, such as flexible printed electrodes around the ear [Debener et al., 2015] or using a limited number of electrodes in the head [Zhang et al., 2022, Nasseri et al., 2020].

2.2.1.2 iEEG

EEG can also be recorded invasively, which is particularly useful to determine the focus hemisphere in TLE or in frontal lobe epilepsy, where the spread of abnormal discharges from one frontal lobe to the other occurs rapidly. Invasive EEG is performed using i) intracranial electrodes that record information directly from the brain or ii) subscalp electrodes, which are subcutaneously implanted between the scalp and the bone [Duun-Henriksen et al., 2020] (see Figure 2.5).

There are three types of intracranial electrodes: subdural strips, subdural grids, and depth electrodes. These are useful to study seizure onset and propagation, where they can be used independently or in combination. Subdural electrodes (grids and strips) require a craniotomy, which is the surgical removal of part of the bone from the skull to expose the brain, to be placed over the cortex surface and, therefore, allow the recording of the Electrocorticographic (ECoG).

With depth electrodes, one can record the brain activity from deeper brain structures, such as the hippocampus, amygdala, orbitofrontal and medial occipital regions. These are implanted through stereotactic surgery, a form of surgical intervention that uses a three-dimensional coordinate system to locate a small target inside the body [Osorio et al., 2016, Jayakar et al., 2016, Spencer et al., 2015]. This signal acquisition is named as Stereoelectroencephalography (SEEG). Since these methods are invasive, the associated risks are higher, where the two major ones are haemorrhage (prevalence rate of 4% per electrode) and infection (infection rate of 3%) [Taussig et al., 2015, Noachtar and Rémi, 2009].

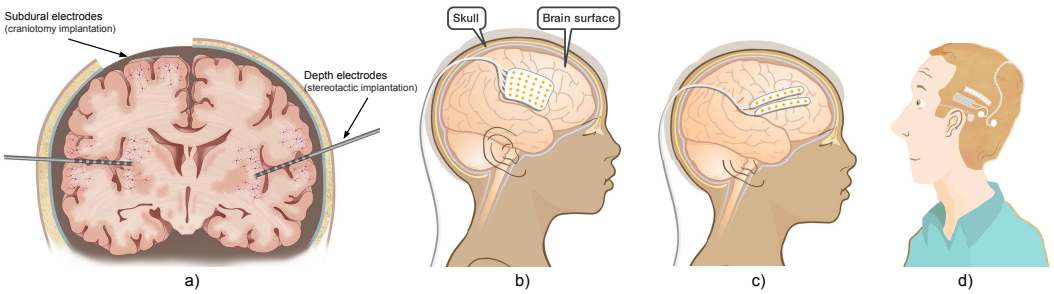


Figure 2.5: Several invasive EEG electrodes: examples of electrode placement for subdural and depth electrodes (a), a subdural electrode grid (b) and a strip one (c), and subscalp electrodes (d). Adapted from [Duun-Henriksen et al., 2020, Grande et al., 2020, Health, 2017].

When comparing intracranial to scalp EEG, the signal-to-noise ratio of the former is higher as the number of intermediate layers between the neural tissue and electrodes is reduced. This difference is reflected in the number of muscle artefacts which is reduced when using iEEG. However, it may be more difficult and not ethical to place intracranial electrodes in a broader brain region than strictly necessary. Thus, despite presenting a higher spatial resolution than the scalp EEG, iEEG may not capture large-scale synchrony as the case of alpha-band activity [Noachtar and Rémi, 2009, Spencer et al., 2015].

Subscalp electrodes comprise a new generation of electrodes as these have the potential to advance treatment, diagnosis, and management of epilepsy for long periods with lesser associated risks. These are implanted subcutaneously via minimally invasive surgery. The subscalp placement removes the need for electrode care, avoids skin abrasions, and secures a stable and low-impedance recording. Compared to standard scalp EEG, subscalp EEG has a reduced spatial resolution, which an result in lower sensitivity for focal abnormalities [Duun-Henriksen et al., 2020, Weisdorf et al., 2019].

2.3 Treatment and therapeutics

The first line of treatment for patients with epilepsy is the administration of AEDs. When medication fails in achieving seizure freedom, other options are possible, such as resective surgery, neurostimulation, dietary therapies, and warning devices. These are essential as patients' continuous exposure to seizures can significantly reduce their quality of life. Patients with uncontrolled seizures tend to present neurologic impairment, namely memory loss, anxiety, and depression [Engel, 2016]. Additionally, due to the unpredictability of seizures, patients might suffer severe injuries related to falls and loss of consciousness, which reinforces social stigma [Devinsky et al., 2018, Laxer et al., 2014].

2.3.1 Antiepileptic drugs and drug-resistant epilepsy

AEDs' goal is to change the balance between excitation and inhibition that characterises epilepsy, which results from hyper excitatory or hypersynchronous neuronal activity. Currently, about 30 AEDs attempt to control seizures by dampening excitatory mechanisms, modulating voltage-gated ion channels, or boosting inhibitory mechanisms [Wang and Chen, 2019].

Currently, medication is not effective for about one-third of patients. When drug administration is unsuccessful, patients are diagnosed with Drug-Resistant Epilepsy (DRE). In Box 4, one can see the formal diagnostic criteria, defined by the ILAE. It is important to note that seizure freedom consists in ceasing all types of seizures for one year or three times the interseizure interval observed before treatment [Kwan et al.,

2010]. Although new AEDs have been developed over the years, the proportion of patients with DRE remained similar. The side effects provoked by medication have decreased [Wang and Chen, 2019, Engel, 2016].

Most patients with DRE are not referred to epilepsy centres where a multidisciplinary team could evaluate them. For instance, less than 1% of patients are referred to an epilepsy centre in the United States of America. Furthermore, for those who are, there is an average gap of 20 years between seizures' onset and referral date, which increases the probability of irreversible damage. It is worth stressing that the mortality rate for DRE is 5-10 times higher than the one of the general population [Engel, 2016].

Box 4 - Diagnostic criteria for DRE (as defined by the Task Force of the ILAE in 2010 [Kwan et al., 2010]).

"Drug-resistant epilepsy may be defined as failure of adequate trials of two tolerated and appropriately chosen and used antiepileptic drugs schedules (whether as monotherapies or in combination) to achieve sustained seizure freedom."

2.3.2 Surgery

When medication fails, resective surgery may be the option for achieving seizure control. Surgery is the most effective way to control seizures in drug-resistant focal epilepsies by resecting the portion of the brain responsible for generating seizures, also known as the epileptogenic zone. Not all patients may undergo this surgery. Its success depends on identifying the epileptogenic zone, which must be limited to a small area. In order to evaluate the possibility of performing the resective surgery, patients undergo presurgical monitoring [Engel Jr, 2018, Engel Jr, 2015, Ryvlin et al., 2014].

During presurgical monitoring, patients are submitted to AEDs withdrawal and sleep deprivation to increase seizure occurrence and to reduce the hospital stay duration [Devinsky et al., 2018, Engel Jr, 2018], where clinicians expect to observe seizures with the same onset characteristics. Nevertheless, there may be some drawbacks since AEDs tapering may trigger generalised tonic-clonic seizures in patients that previously did not experience seizures of this type. It also may increase the risk of having seizure clusters that lead to severe damage. Due to this, sleep deprivation and medication are introduced with care and using established protocols [Kirby et al., 2020, Rathore and Radhakrishnan, 2015].

In order to localise and delineate the epileptogenic zone, clinicians use a multimodal approach: long-term video-EEG, structural Magnetic Resonance Imaging (MRI), and a neuropsychological evaluation. With this information, patients undergo resective surgery if: i) the different approaches present coherent findings, ii) there is a well-defined epileptic region, and iii) there is a reasonable risk-benefit

ratio [Engel Jr, 2018, Rathore and Radhakrishnan, 2015]. Figure 2.6 depicts this process.

When this process fails to identify or delineate the epileptogenic region, other signals can be acquired, such as Magnetic Source Imaging (MSI), functional MRI, Single-Photon Emission Computed Tomography (SPECT), and Positron Emission Tomography (PET). With these, clinicians verify if there is a chance of generating a testable hypothesis regarding the epileptogenic zone. In a positive case, the patient will undergo intracranial EEG acquisition, cortical stimulation, and mapping. If the epileptogenic zone can be localised and resected, the patient will undergo surgery [Rathore and Radhakrishnan, 2015].

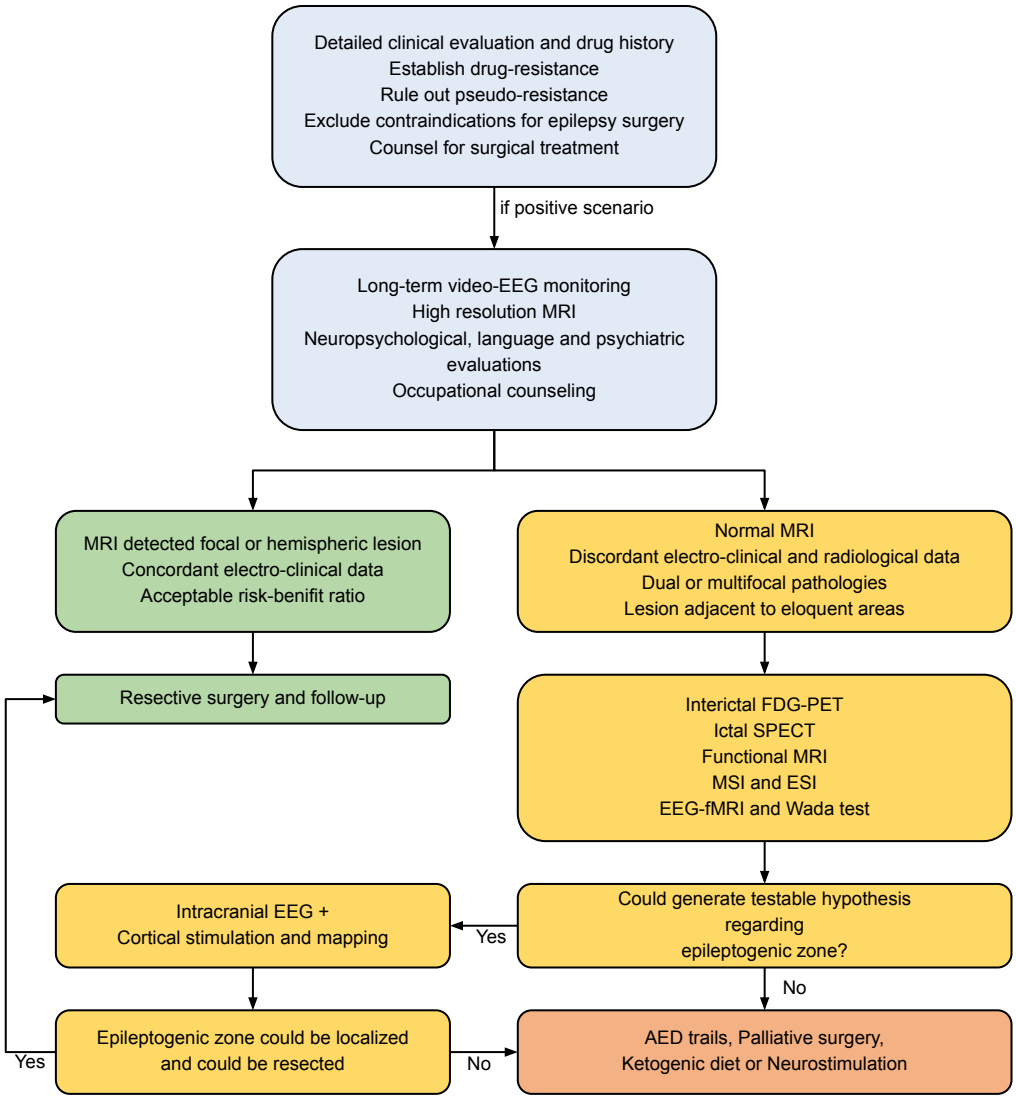


Figure 2.6: Flow chart depicting the usual process of presurgical evaluation and selection for surgery. Blue boxes are common processes for all patients. Green represents the best case scenario, yellow the intermediate scenario, and orange the worst one. Adapted from [Rathore and Radhakrishnan, 2015].

After surgery, in patients with TLE, the seizure-free rate within 10 years varies among studies [Ryvlin et al., 2014], from 49% [De Tisi et al., 2011] to 83% [Murphy et al., 2010] of patients achieving seizure-freedom. Advances in neuroimaging techniques lead to an improvement in this treatment option. The eligibility criteria have widened, and surgery is now available for more patients. It is also important to note that many patients decline presurgical monitoring. These refusals may be due to misconceptions and fears both from clinicians and patients [Engel Jr, 2018, Ryvlin et al., 2014].

2.3.3 Neurostimulation

Neurostimulation can be offered after confirming a patient's ineligibility to undergo resective surgery. It consists in implanting a device that delivers electrical pulses to peripheral nerves or specific brain areas of the central nervous systems to prevent potential seizures. This option is palliative as only a reduced portion of patients become seizure-free for more than one year [Ryvlin et al., 2021, Rincon et al., 2021, Bigelow and Kouzani, 2019, Boon et al., 2018, Krishna et al., 2016].

Current neurostimulation techniques can be divided into invasive and noninvasive, depending on the need to perform surgery for implantation. They can also be categorised into open and closed-loop when a scheduled or responsive intervention is considered, respectively. Invasive strategies are the most used, and include Vagus Nerve Stimulation (VNS), Deep Brain Stimulation (DBS), and Responsive Neurostimulation System (RNS). Noninvasive options concern transcutaneous vagus nerve stimulation, trigeminal nerve stimulation, and transcranial magnetic stimulation [Ryvlin et al., 2021, Rincon et al., 2021, Bigelow and Kouzani, 2019, Boon et al., 2018]. However, these are not clinically validated. To better understand how VNS, DBS, and RNS strategies work, see Figure 2.7, which provides an intuitive vision of the neurostimulation brain targets and primary anatomical pathways, and Table 2.1, which presents a comparison between the three solutions.

VNS was the first approved neuromodulation therapy. It started as an open-loop strategy, stimulating the left vagus nerve within a defined rhythm, typically 30 seconds every 5 minutes. Now, it works as a closed-loop system that delivers stimulus upon predefined heart rate change patterns, believed to trigger a seizure generation process [Ryvlin et al., 2021]. VNS procedure is extracranial; therefore, its surgery is less risky, with tolerable side effects. Also, it is possible to use a specific magnet to either stop the device or deliver a single stimulation [Ryvlin et al., 2021, Ryvlin et al., 2014, Laxer et al., 2014].

DBS delivers a scheduled stimulus of one minute every five minutes to the Anterior Nucleus of the Thalamus (ANT), assumed to contribute to seizure generation significantly. Compared to VNS, its procedure involves more risks as it requires brain surgery to place intracranial electrodes and chest surgery to insert a pulse

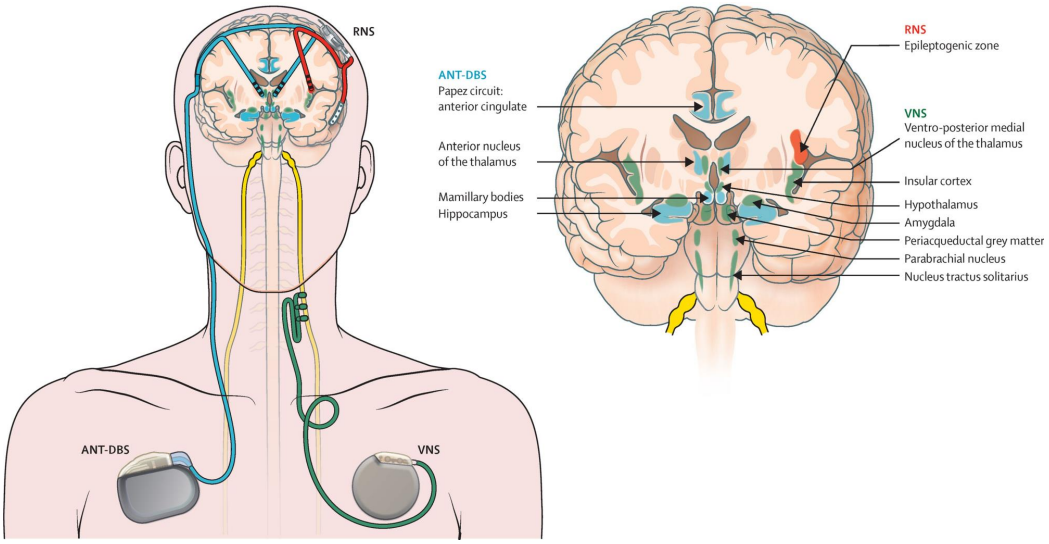


Figure 2.7: Approved neurostimulation therapies for DRE, also showing the brain targets for each neuromodulation approach according to sites of stimulation and known primary anatomical pathways. Source: [Ryvlin et al., 2021].

generator [Rincon et al., 2021, Laxer et al., 2014]. Although exact action mechanisms are not fully known, high-frequency stimulation can regulate abnormal brain impulses and disrupt epileptic networks [Rincon et al., 2021, Bigelow and Kouzani, 2019, Markert and Fisher, 2019].

Table 2.1: Approved neuromodulation therapies for epilepsy with implantable devices [Ryvlin et al., 2021, Bigelow and Kouzani, 2019, Schulze-Bonhage, 2019].

	VNS	**ANT-DBS**	**RNS**
Target	Left vagus nerve	Anterior nucleus of the thalamus	Ictal onset zone (cortex)
Stimulation type	Open-loop; or heart rate responsive closed-loop; or on demand	Open-loop	Ictal intracerebral EEG responsive closed loop
Data collected	Therapy activations; prone position; heart rate	Local field potentials via live streaming in clinic	Seizure frequency; EEG ictal discharge
Age	Children$\geq$ 4 years and adults	Adults	Adults
Type of seizures	Focal and generalised	Focal	Focal
Epileptogenic focus or foci	Non-localisable, multifocal, or not resectable	Bitemporal, multifocal, or non-localisable	Bitemporal or eloquent focus
Short-term infection rate	1% at 3 months	Not available	3% at 3 months
Long-term infection rate	Not available	13% at 10 years	12% at 9 years
Material dysfunction	1%	8% lead replacement	5% lead damage or revision
Other frequent adverse events during first year	29% hoarseness; 12% paraesthesia; 8% shortness of breath; 8% cough	18% paraesthesia; 11% implant site pain; 15% depression; 13% memory complaints; 4% headache	9% dysesthesia or paraesthesia; 16% implant site pain; 3% depression; 4% memory complaints; 11% headache
50% responder rate at 1 year	37%	43%	44%
Seizure free at 2 years of follow up	0%	13%	9%

RNS is a closed-loop system that stimulates the cortex directly in up to two epileptogenic regions. It is necessary to implant a neurostimulator along with depth electrodes or cortical strip leads, able to read EEG to detect seizure generation activity patterns through a threshold-based model. EEG features capture changes in amplitude, frequency, and rhythmicity [Ryvlin et al., 2021, Laxer et al., 2014, Sun and Morrell, 2014].

Although Table 2.1 presents a comparison between VNS, ANT-DBS, and RNS, there are no studies that compare these techniques directly in a randomised form. Several parameters may significantly difficult the comparison of strategies, as in the case of trial design, stimulation parameters, and medication. Despite there may be a lack of consensus across epilepsy centres on which strategy to adopt and how [Ryvlin et al., 2021, Rincon et al., 2021, Markert and Fisher, 2019, Schulze-Bonhage, 2019, Boon et al., 2018], the literature suggests a higher performance in ANT-DBS and RNS when compared to VNS. The VNS is the most available treatment as it is considerably less invasive and thus provides the most favourable risk/benefit ratio [Ryvlin et al., 2021, Laxer et al., 2014].

The responder rate is a common measure of epilepsy therapy efficacy, defined as the percentage of patients achieving $\geq$50% reduction in seizure frequency during a specified period. At one year, a responder rate of 37%, 43% and 44% were reported for VNS, ANT-DBS, and RNS, respectively. Other efficacy values can be found in the literature, and a substantial part may present a high risk of bias [Ryvlin et al., 2021]. In addition to the presented risks and adverse events presented in Table 2.1, all these options need a battery replacement, resulting in frequent clinical appointments to adjust parameters [Ryvlin et al., 2021]. Several concerns have also arisen in the closed-loop response of the RNS system [Sun and Morrell, 2014], as there are reports of an overresponsive behaviour where the majority of interventions may be unnecessary. It may be possible that the overresponsiveness of VNS and RNS may arise from frequent stimulation that drives long-term neuromodulation rather than from detecting pre-seizure activity [Schulze-Bonhage, 2019, Markert and Fisher, 2019].

2.3.4 Rescue medication

Rescue medication also plays a vital role in epilepsy as it may: i) provide seizure freedom when combined with AEDs; ii) reduce the continuous intake dose of AEDs and, therefore, reduce its long-term side effects; and iii) stop seizure clusters and prolonged seizures [Beniczky et al., 2021, Baud et al., 2020, Gaínza-Lein et al., 2017]. The drugs used as rescue medication for epilepsy are benzodiazepines due to their rapid effect. These should be used as an acute treatment due to their long-term adverse secondary effects, strong addiction, and habituation [Riss et al., 2008].

Diazepam rectal gel was the first and only benzodiazepine approved by the Food

and Drug Administration (FDA), for several years, as a rescue medication for seizure cluster treatment outside the hospital, mainly for children's use. Currently, intranasal and buccal routes have been developed since they are more accessible and, thus, less invasive. Midazolam nasal spray is used in patients older than 11 years, and diazepam nasal spray for patients older than six. In Europe, Buccal midazolam was approved for treating prolonged seizures in children and teenagers [Cloyd et al., 2021, Bouw et al., 2021]. Table 2.2 shows the different options for emergency medication, along with its route, peak effect level, and time to take effect.

2.3.5 Warning devices

Intervention devices have been explored for epilepsy management, particularly for seizure detection. The ability to continuously monitor a biosignal and to swiftly detect/predict a seizure, followed by raising an alarm, may provide the patient or caregiver enough time to minimise seizure consequences or to allow the intake of rescue medication [Beniczky et al., 2021, Nasseri et al., 2020]. These are designed to integrate algorithms that analyse long-term signals and generate alarms while excluding data segments containing artefacts. Several new EEG acquisition systems are now available including UNEEG SubQ, EpiMinder Subscalp, and Byteflies Sensor Dots [Baud et al., 2022]. Besides the EEG signal, researchers also consider other noninvasive signals due to patient comfort, such as accelerometry, electrodermal activity, photoplethysmography, electromyography, body temperature, and electrocardiography [Nasseri et al., 2020, Gadhoumi et al., 2016b, Ramgopal et al., 2014]. This is now possible due to devices [Brinkmann et al., 2022] such as Empatica E4, Fitbit Charge HR, and Fitbit Inspire, for example.

The NeuroVista Seizure Advisory System [Cook et al., 2013] is one of the most relevant warning devices (NCT01043406) developed for seizure prediction. A phase I clinical trial was performed on 15 patients with medically refractory epilepsy for two years. This system used intracranial electrodes that recorded the EEG to monitor

Table 2.2: Rescue medication options approved for treatment outside the hospital [Cloyd et al., 2021, Bouw et al., 2021, Wolf et al., 2012, Boddu and Kumari, 2020, Boddu and Kumari, 2020].

Drug	Route	Time to take effect	Peak level	Approval
Diazepam	Rectal	5-10 minutes	10-45 minutes	FDA in 1997 for seizure clusters
Midazolam	Buccal	<5 minutes	20-30 minutes	European Union in 2011 for prolonged seizures in non-adults
Midazolam	Intranasal	<10 minutes	15-120 minutes	FDA in 2019 for seizure clusters in patients older than 11 years
Diazepam	Intranasal	<5 minutes	>60 minutes	FDA in 2020 for seizure clusters in patients older than 5 years

the brain continuously and gave the patients a likelihood of seizure occurrence. The advisory system comprises three main components: leads of electrodes, an implantable telemetry unit, and the personal advisory pad. The leads were placed according to the patients' epileptogenic zone. Then, they were tunnelled down the neck to the telemetry unit and implanted in the chest. This telemetry unit transmitted wireless information to the pad like a pager.

This study was the first to demonstrate a real possibility for prospective seizure prediction using ambulatory EEG data. Firstly, there was a model training phase, where only the best-performing patients continued to the advisory phase. Periodic retraining was found necessary over time to maintain or improve performance. The patients' response to the advisory system was not homogeneous, which could be explained by the high variability of seizure warning times. Patients with the lowest proportions of time in high seizure likelihood claimed to be satisfied and benefited from it. Another interesting finding of this study was that the seizure events reported by patients significantly differed from the ones captured by the EEG, which may raise doubts concerning the use of patient diaries as a gold standard for evaluating the success of any therapeutic procedure in epilepsy [Gadhoumi et al., 2016b].

Lastly and very importantly, the advisory system used Machine Learning (ML) models that were not entirely transparent and, thus, black boxes. The fact that the authors have used black-box models in the clinical trial is a critical topic that this thesis will deeply address in chapter 4.

2.4 Seizure Prediction

Seizure prediction is the focus of this thesis. This scientific area aims to build a tool to read online data and timely inform the patient about an upcoming seizure occurring on a well-defined future time window after a specific horizon [Assi et al., 2017, Osorio et al., 2016, Gadhoumi et al., 2016a, Mormann et al., 2007].

It is possible to divide the EEG from a patient with epilepsy in different periods in time, as seen in Figure 2.8: a preictal period, which precedes the seizure; the ictal period corresponding to the seizure; the postictal period, which follows the seizure; and finally, the interictal period, found in between the postictal and the preictal of consecutive seizures.

In seizure prediction, it is vital to differentiate the interictal period from the preictal one, which relies on correctly identifying seizure biomarkers that capture the transition from a seizure-free state to a seizure. The preictal period is the most challenging interval to annotate as there is no recurrent pattern, which concerns the primary difficulty in seizure prediction. It is associated with significant heterogeneity from patient to patient and from seizure to seizure. For this reason, patient-specific models have proven to be more successful than general models [Kuhlmann et al., 2018b, Freestone et al., 2017, Mormann et al., 2007].

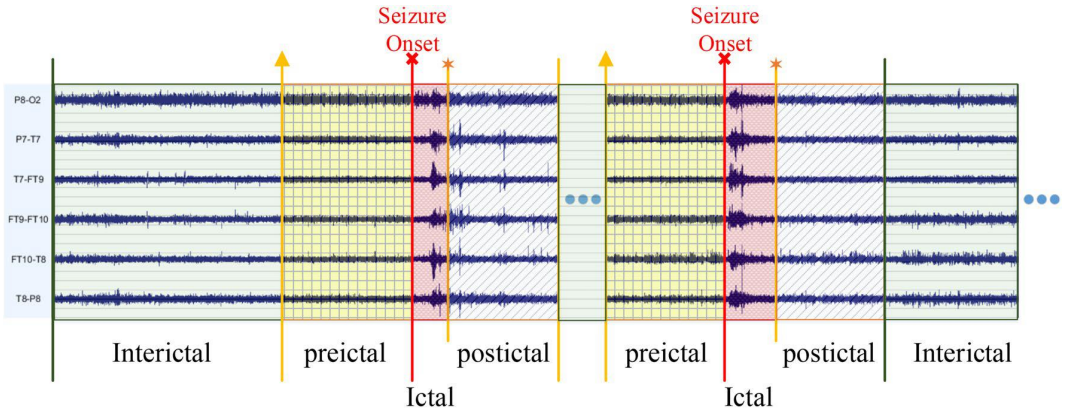


Figure 2.8: Example of an EEG signal where it is possible to visualise the different brain states (interictal, preictal, ictal, and postictal) concerning a seizure. Source: [Cui et al., 2018].

2.4.1 Seizure onset

It is fundamental to obtain, along with the EEG, annotations of the seizures' onset. In presurgical monitoring conditions, these are obtained with video-EEG. There are two onsets: i) the EEG onset, which concerns the moment where significant changes in the EEG are observed; and ii) the clinical onset, which concerns the appearance of the first symptoms, derived from the neurophysiological alterations. Prediction algorithms usually consider the EEG onset, as these models are based on the EEG signal. The clinical onset may not always be identified [Osorio et al., 2016, Varsavsky et al., 2011, Mormann et al., 2007].

2.4.2 Lead seizure

Authors need to analyse a considerable number seizures to obtain confidence in the prediction results. It is also necessary to handle seizure clusters, as the prediction models are developed by considering seizures as independent events. Authors usually only try to predict the first seizure of a seizure cluster, called a lead seizure [Chen and Cherkassky, 2020]. As there is no consensus regarding the seizure cluster definition, authors tend to choose a minimum seizure-free interval to claim that the further seizure is an independent event. The selection of this interval usually concerns the need for having a reasonable number of seizures. Studies have considered lead seizures as consecutive seizures separated by, at least, 1 hour [Stirling et al., 2021b], 1.5 hours [Alvarado-Rojas et al., 2014], 2 hours [Meisel et al., 2020, Meisel and Bailey, 2019], 4 hours [Nasseri et al., 2021], 4.5 hours [Lopes et al., 2021], 5 hours [Karoly et al., 2017], and 8 hours [Cook et al., 2013].

2.4.3 Detection vs prediction

A similar research field is seizure detection, where authors may attempt to inform about the occurrence of a seizure in the very last moments of preceding the seizure (in the order of seconds), which still allows a swift intervention, such as neurostimulation triggered through closed-loop systems. Seizure prediction, however, offers a more extended period for preparing for the seizure and minimising consequences (from minutes to hours). Authors also use seizure detection for other applications as it has proven helpful in identifying the seizure focus and performing a semi-automatic inspection of seizures' EEG-onset under clinicians' supervision [Bialer et al., 2017, Ramgopal et al., 2014, Mormann et al., 2007].

2.4.4 Forecasting vs prediction

Seizure forecasting is a parallel area that gained significant interest. When compared to seizure prediction, it shifts away from whether a seizure occurs or not (preictal state) to identify periods of a high probability of seizure occurrence (proictal state) [Baud et al., 2020, Dumanis et al., 2017]. In other words, while prediction develops classifiers capable of distinguishing interictal/preictal chunks of EEG data over time, forecasting relates to a probabilistic view, where a momentary seizure risk can be assessed by tracking cycles related to EEG activity. These cycles may concern probing cortical excitability [Dell et al., 2019] or interictal epileptiform activity (epileptic spikes, high-frequency oscillations, or rhythmic bursts of high-amplitude oscillations) [Proix et al., 2021, Maturana et al., 2020, Baud et al., 2018].

One of the significant limitations of prediction and other EEG epilepsy-related research area is its focus on fast aspects ast aspects of the EEG time series at the expense of a poor understanding of slow variable keys that might also explain a transition to an ictal state. It is worth noting that this tendency might be explained by the excellent temporal resolution of the EEG signal. Thus, besides the paradigm change to a proictal state, forecasting also shifts its time horizon from minutes to days. Although Concept Drift (CD) may influence prediction success, they may play a more prominent role in forecasting since the proictal state might be forecasted on a scale of hours to days by tracking multiday and circadian cycles while accounting for seizure frequency and sleep dynamics [Proix et al., 2021, Maturana et al., 2020, Baud et al., 2018]. CDs are deepened in Background section 2.5 and in the State of the art section 3.2.

2.4.5 Seizure prediction characteristic

In the early 2000s, several methodologies had already been proposed for EEG seizure prediction, but proper evaluation and comparison were hard as no recognised criteria existed. It was also challenging to assess whether the performance of an algorithm was sufficient for clinical application. Summarily, a prediction method

analyses chronological windows of EEG, where it timely raises alarms that allow an intervention.

Based on this, [Winterhalder et al., 2003] proposed in 2003 the seizure prediction characteristic, which is a framework to evaluate seizure prediction methods based on clinical, behavioural, and statistical considerations. The evaluation was based on two metrics adapted to an alarm system: seizure sensitivity and False Prediction Rate per hour (FPR/h). Furthermore, it also accounts for other two concepts: Seizure Prediction Horizon (SPH) and Seizure Occurrence Period (SOP). SPH is the time interval that a prediction tool guarantees for the patient to prepare for the forthcoming seizure, constituting a zero probability of seizure occurrence. In other words, it corresponds to the period between the fired alarm and the beginning of SOP. It is also known as Intervention Time (IT). SOP is the period where a seizure is expected to occur. Thus, an alarm is considered correct (a true positive) when the seizure occurs during the SOP period. Suppose an alarm is fired and no seizure occurs within the considered SOP. In that case, the alarm is considered false (a false positive) (see Figure 2.9 for an example).

The SOPs found in the literature range from minutes to hours [Mormann et al., 2007], which constitutes a sensitive matter as it can significantly impact the performance of the algorithms in the patient's life and, naturally, in the chosen intervention. If the used SOP is tremendously long, it may not be helpful. For instance, when simply envisioning a warning system, a SOP of eight hours may be useless if a given patient has three seizures per day, as the used algorithm will not technically make any prediction despite accurately anticipating all seizures. Furthermore, adjusting the parameters of a prediction methodology may be arduous, as higher sensitivity values may increase the FPR/h. The selection of acceptable parameters must consider all factors simultaneously. Too many false alarms may lead a patient to ignore the warning system or lead to side effects of unnecessary interventions and, consequently, worsen a helplessness situation [Mormann et al., 2007].

Due to this, the increase of sensitivity is questionable at the expense of a high FPR/h. The same problem happens with SPH and SOP as large intervals may lead

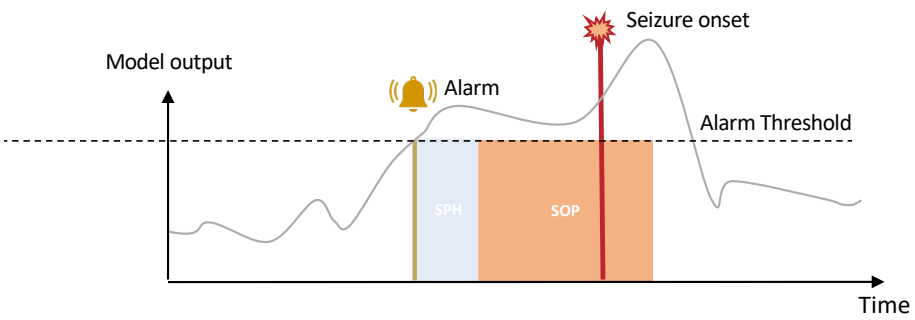


Figure 2.9: An example of a true alarm in a seizure prediction methodology, while presenting the SPH and SOP concepts. Adapted from [Winterhalder et al., 2003].

to psychological stress and anxiety. In contrast, the SPH must provide enough time for intervention and to self-prepare to minimise the consequences of a seizure [Schelter et al., 2007, Mormann et al., 2007, Winterhalder et al., 2003]. Summarily, to allow intervention, a proper prediction methodology must be associated with reasonable SOP and SPH periods. These systems must predict the majority of seizures to provide confidence to the patient and present a low value of false predictions to minimise the effects of too many unnecessary interventions.

2.4.5.1 Performance assessment

Based on the seizure prediction characteristics, two fundamental metrics arise from the standard machine learning sensitivity and specificity measures. If seizure prediction would be a classic machine learning problem, the binary classification of interictal/preictal samples would be evaluated by accessing sample sensitivity (SS_{sample} in Equation 2.1) and sample specificity (SP_{sample} in Equation 2.2), as shown in Figure 2.10. Despite these may not truly access the evaluation of an alarm system, many authors use them, along with Receiver Operating Characteristic (ROC) curve, to solely demonstrate the discriminative potential of developed classifiers to classify independent segments of EEG data [Kuhlmann et al., 2018b, Mormann et al., 2007].

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

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

Seizure sensitivity and FPR/h are adjusted to the context of an alarm system in a time-series signal. Intuitively, one could envision the preictal period as the interval where the anticipating method must raise the alarm. If it triggers an alarm during the interictal period or the SPH period, it is a false alarm [Gadhoumi et al., 2016b, Winterhalder et al., 2003].

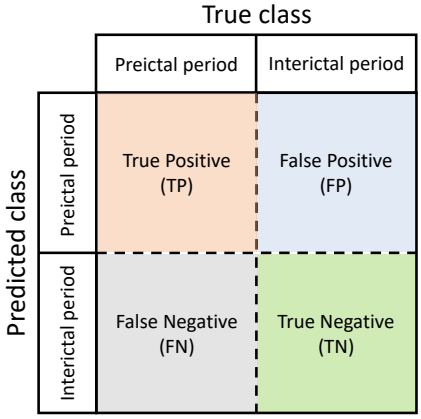


Figure 2.10: Confusion matrix for assessing sample seizure prediction.

Seizure sensitivity (SS in Equation 2.3) is obtained through the ratio between the number of true alarms and total analysed seizures. The FPR/h (Equation 2.4) is calculated by the number of false alarms divided by the period during which the model can raise a false alarm ($\triangle_{interictal}$), namely the interictal period [Winterhalder et al., 2003]. Since each triggered alarm has associated a given SPH and a SOP, no further alarms can be raised during that period. This period is named as refractory period ($\triangle_{SPH} + \triangle_{SOP}$), and needs to be discarded from the FPR/h denominator. This operation also helps to compare models from different methodologies as not all authors use refractory periods in their systems [Mormann et al., 2007]. Figure 2.11 illustrates how these metrics are calculated. Additionally, it is important to note that the FPR/h metric might be particularly useful as it concerns a specificity-related measure that allows the clinician to understand/or study what might be the patient complacency toward false interventions. For example, when developing a warning system, a high rate of false alarms may lead the patient to increased anxiety levels and, inevitably, to lose confidence in the device [Schulze-Bonhage et al., 2010, Schelter et al., 2008]. In the case of a neurostimulation closed-loop system, there may be a higher tolerance toward false alarm interventions [Sun and Morrell, 2014]. Thus, the accepted performance values might change depending on the intervention.

$$SS = \frac{\#true\ alarms}{\#seizures}, \tag{2.3}$$

$$FPR/h = \frac{\#false\ alarms}{\triangle_{interictal} - \#false\ alarms(\triangle_{SPH} + \triangle_{SOP})}. \tag{2.4}$$

Ideally, the performance of a seizure prediction model would be a sensitivity of one (all seizures predicted) and a null FPR/h (no false alarms raised). This scenario has proven difficult over the years, where a trade-off between these metrics has been reported. Again, for the specific case of a warning in presurgical monitoring, [Winterhalder et al., 2003] proposed a maximum FPR/h of 0.15, according to patients' mean seizure occurrence rate in these conditions (3.6 seizures per day). As patients with DRE during regular daily life have a mean seizure rate of three per month, one can compute a maximum FPR/h of 0.0042. Again concerning a warning system, most patients claimed the need for a minimum of a 0.90 sensitivity performance [Schulze-Bonhage et al., 2010].

2.4.5.2 Statistical validation

Another critical aspect of seizure prediction is statistical validation: a developed model must overperform, with statistical significance, a predictor based on chance-level [Assi et al., 2017, Mormann et al., 2007]. The two most widely used approaches concern unspecific predictors [Schelter et al., 2008, Winterhalder et al., 2003], and

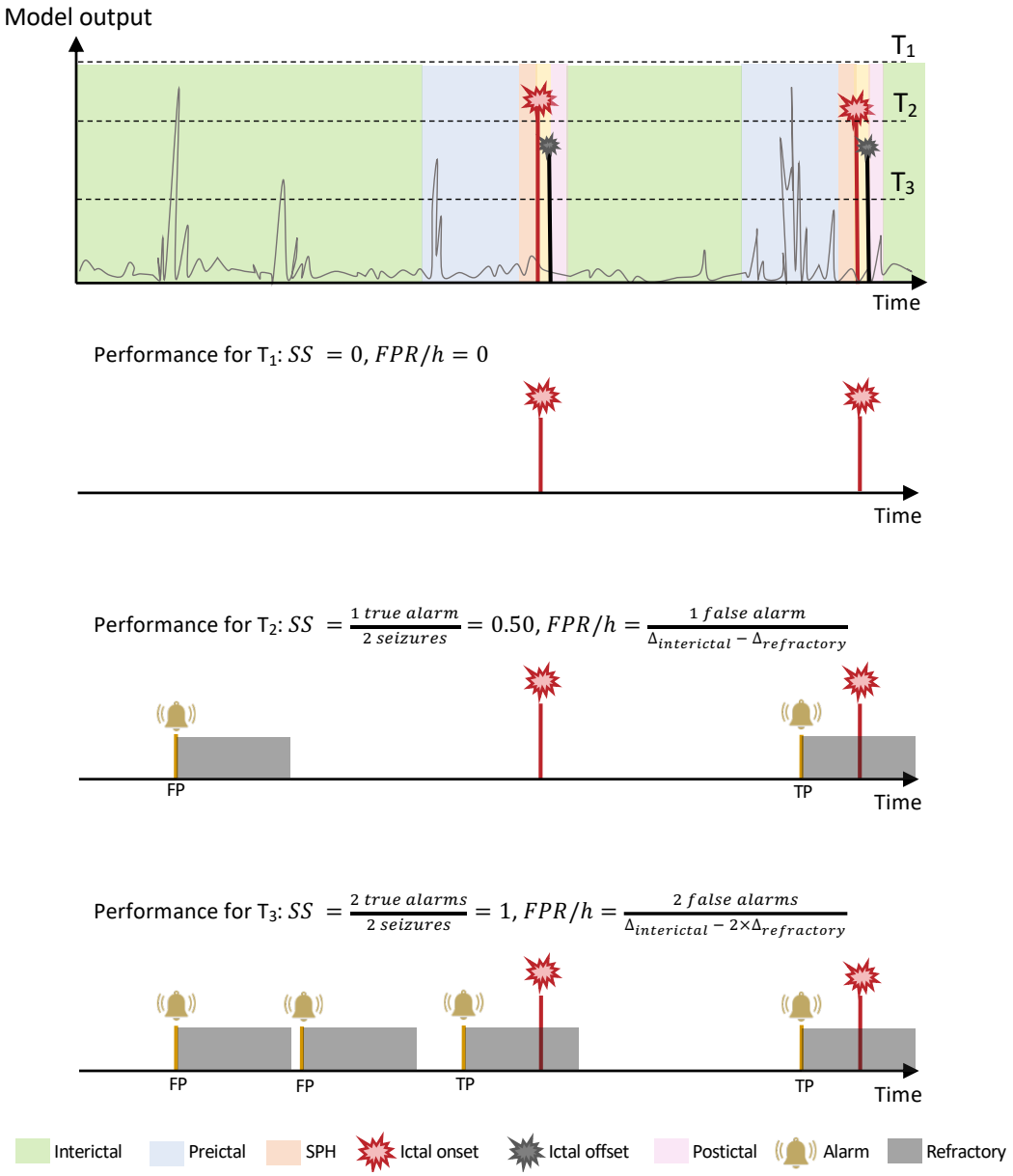


Figure 2.11: Example of the obtained seizure prediction characteristic performances, where alarms are triggered whenever a model output crosses a defined threshold. Three performances are presented, one for each defined threshold (T_1, T_2, and T_3). This example also demonstrates the trade-off between seizure sensitivity and FPR/h. Adapted from [Winterhalder et al., 2003].

the surrogate analysis [Andrzejak et al., 2009, Schelter et al., 2008].

Unspecific predictors

[Winterhalder et al., 2003] proposed a method in which alarms are triggered randomly without using any information from the EEG signal. Firstly, it is possible to calculate the probability of occurring an alarm p during a small interictal time interval I, as described by Equation (2.5):

$$p = FPR/h \times I. \tag{2.5}$$

For a longer time interval W, the probability P of at least one alarm to occur can be calculated with Equation (2.6):

$$P = 1 - (1 - FPR/h \times I)^{W/I} \approx 1 - e^{-FPR/hW} for\ I \ll W. \tag{2.6}$$

For $W = SOP$, it represents the sensitivity of a random prediction method as it constitutes the probability of at least one alarm being raised during the seizure occurrence period. FPR/h values over 0.15 [Winterhalder et al., 2003] are not ready for clinical application.

Later, [Schelter et al., 2008] also proposed an unspecific analytic predictor, known as the random predictor, which is based on a homogenous Poisson process for false predictions. At any single sampling point of a feature extracted from a time series, the probability of raising an alarm is given by Equation (2.7):

$$P_{Poiss} = \frac{FP}{N}, \tag{2.7}$$

where FP is the number of false alarms, and N is the number of samples. When considering a time period with equal duration to SOP and that the product of FPR/h with SOP is considerably smaller than one, the probability P from (2.6) of raising at least one alarm within SOP can be approximated to Equation (2.8):

$$P \approx 1 - e^{-FPR/hW} \approx FPR/h \times SOP. \tag{2.8}$$

The previous equation assumes that the patient is not under continuous warning. The probability P forms the basis for a significance level that will assess if the sensitivity, $SS(FPR/h, SOP)$, of a prediction algorithm under test outperforms the one from a random predictor.

Furthermore, this statistical method considers the analysis of more than one seizure, where it also increases random prediction performance by using more electrodes (d). The latter is due to the false prediction rate, which is usually not the same for all channels. Therefore, including more channels and/or measures leads to increased seizure prediction by chance. When using ML models, there is only one prediction output and one FPR/h, as the correspondent methods can use multidimensional inputs to obtain a single-dimension output (d=1). In this context, d >1 is used when several predictions are run simultaneously. Thus, the probability of predicting at least k of K seizures is described in the form of an cumulative binomial distribution, as in Equation (2.9):

$$P_{binom}(k, K, P) = 1 - \left[\sum_{1}^{j \leq k} \binom{K}{j} P^j (1 - P)^{K-j} \right]^d. \tag{2.9}$$

Then, the critical value to test statistical significance is calculated with Equation (2.10):

$$\sigma = \frac{argmax_k\{P_{binom}(k, K, P) > \alpha\}}{K} \times 100\%. \tag{2.10}$$

In sum, the random predictor's advantage lies in its analytic expression, which does not require the EEG signal and is therefore computationally light. Nevertheless, it is based on a homogeneous Poisson process and thus, assumes a homogeneous distribution of false alarms over time, which may not allow dealing with some seizure dynamics such as seizure non-random occurrence induced by CDs or medication withdrawal. Additionally, beating the random predictor may be significantly challenging for a low number of tested seizures, which may be a problem as seizures are rare events and methodologies tend to be patient-specific. To better understand the random predictor behaviour, Figure 2.12 shows the variation of its sensitivity regarding SOP duration and number of tested seizures k.

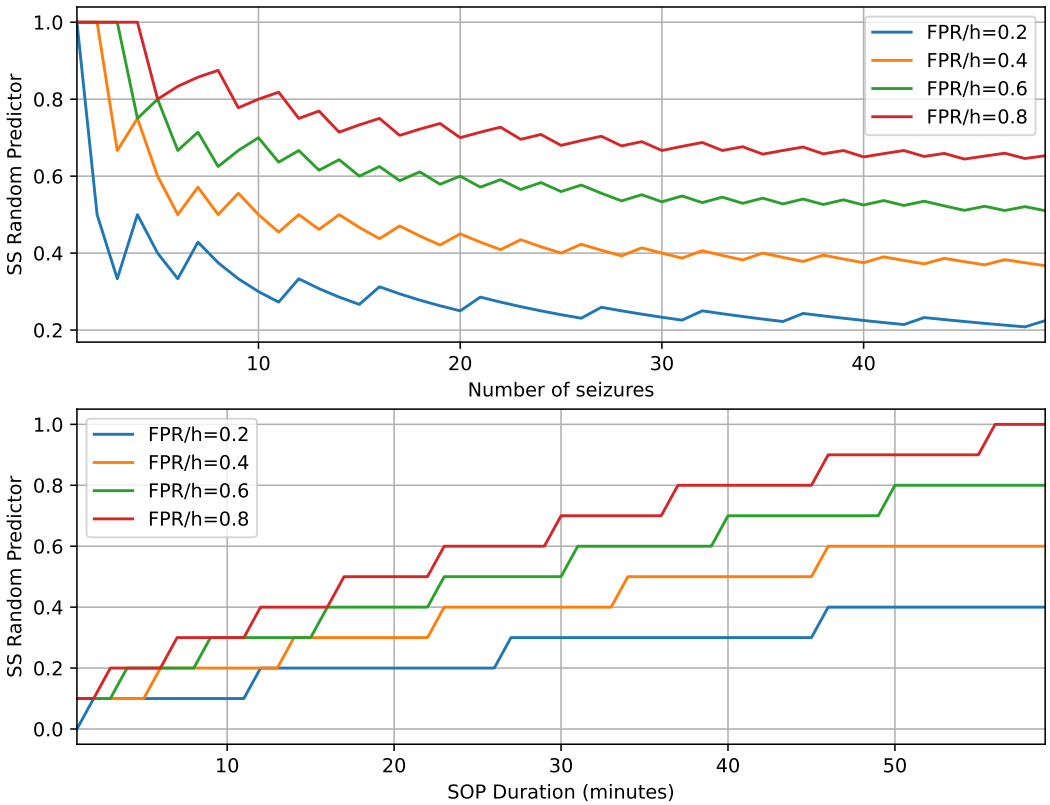


Figure 2.12: Example of the random predictor sensitivity behaviour, towards an increase in the number of seizures (above) and SOP duration (below). For (above), a SOP of 40 minutes was used, and for (below) $k = 10$ seizures were used. In both cases, a significance *alpha* of 0.05 was used.

Surrogate analysis

Surrogate time series analysis is a strategy based on Monte Carlo simulations developed from constrained randomisations of the original seizure times, sharing specified properties of the original data. Then, the surrogate performance is assessed and compared with the developed method's performance. If the developed method outperforms the surrogate one with statistical significance, it is possible to claim that the new method performs better than chance. Surrogate-based techniques have the advantage of being more flexible, particularly when the assumptions made by the random predictor are not reasonable, such as the false alarm occurrence following a Poisson distribution [Schelter et al., 2008, Mormann et al., 2007]. When using these strategies, it is essential to pay special attention to their implementation, as they allow the testing of different null hypotheses [Andrzejak et al., 2009, Andrzejak et al., 2003].

There are two major strategies: seizure-times surrogates [Andrzejak et al., 2003] and measure-profile surrogates [Kreuz et al., 2004]. In the first, original seizure times are replaced with random times, assuming the maintenance of the inter-seizure-interval distribution (random permutation of inter-seizure-intervals). In the second, original measure profiles are randomised so that no feature is related to the true predictive power of the seizure predictor, which can be performed using the simulated annealing technique. A significant drawback of this approach is its demanding computational power [Schelter et al., 2008]. For this thesis, a surrogate time-series analysis was the chosen strategy for statistical validation, as depicted in Figure 2.13.

2.4.5.3 Postprocessing

Since the developed classifiers are trained to make classifications on independent EEG segments, there is a need to handle the temporal relations between each classifier output. It would also be unrealistic to consider the raw output of the classifier as a fair alarm generator: is it unlikely to classify all samples correctly, and it is very common to encounter noise in online data, particularly during long-term recordings [Assi et al., 2017, Teixeira et al., 2014b]. In order to prevent these problems, postprocessing methods are often applied, such as the Kalman filter [Park et al., 2011, Chisci et al., 2010] and the Firing Power [Teixeira et al., 2012].

The Kalman filter underlying idea consists in the estimation of the states s_k of a linear dynamic system at instant k, where y_k denotes the measured variable, and w_k and z_k are zero-mean white noise vectors (Equation (2.11)). An alarm is raised whenever the Kalman filter output is classified as a preictal sample. Also, new alarms can only be raised when the output crosses the zero-threshold in an ascending way [Teixeira et al., 2012].

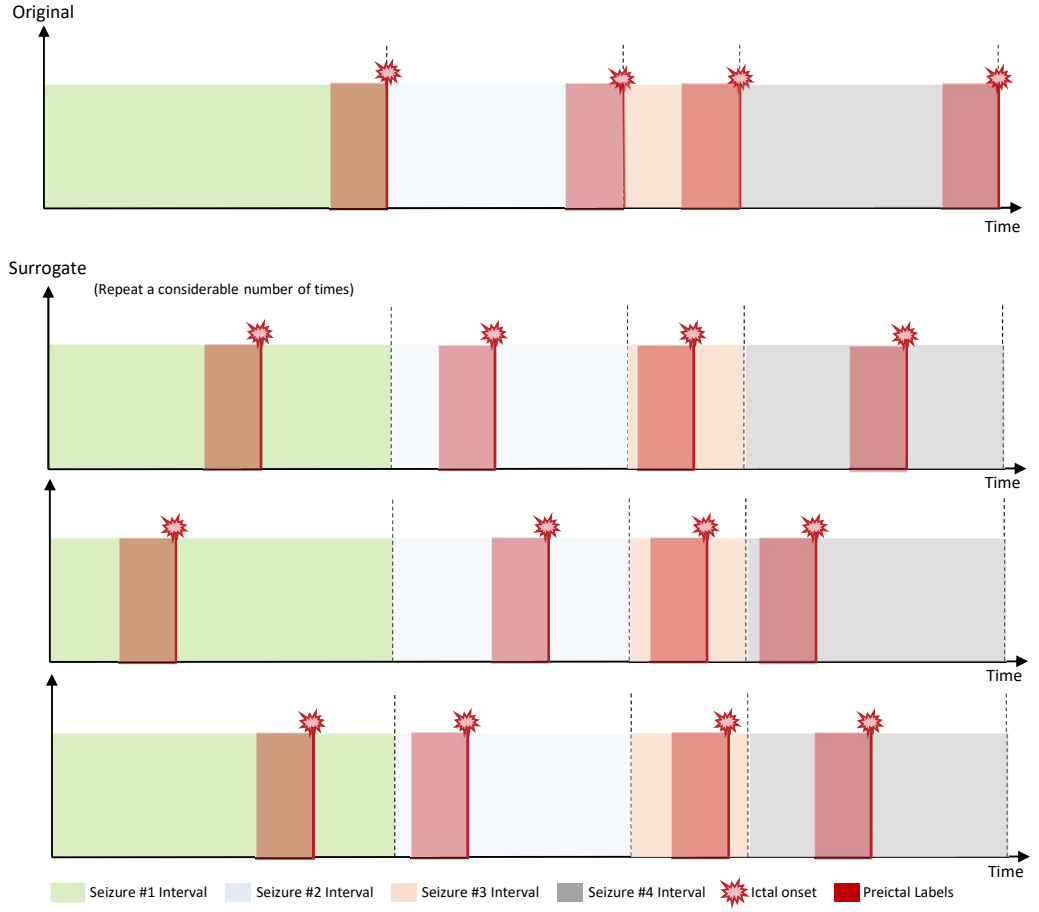


Figure 2.13: Example of the used seizure-times surrogate analysis. Original seizure times and random permutation of inter-seizure intervals in such a way to maintain its seizure occurrence frequency. Adapted from [Schelter et al., 2008].

$$\begin{cases} s_{k+1} = \begin{bmatrix} 1 & T_p \\ 0 & 1 \end{bmatrix} s_k + w_k \\ y_k = \begin{bmatrix} 1 & 0 \end{bmatrix} s_k + z_k \end{cases} . \tag{2.11}$$

The Firing Power method [Direito et al., 2017, Teixeira et al., 2014b] (see Figure 2.14) applies a moving average filter to the output of the binary classifier, where its window size is the same as the SOP duration. In other words, the Firing Power provides, over time, a measure of the ratio of samples classified as preictal within a window with a duration equal to the considered preictal period. It is possible to mathematically describe this moving average filter as shown in Equation 2.12, where τ is the filter window, $fp[n]$ the Firing Power output at instant n:

$$fp\,[n] = \frac{\sum_{k=n-\tau}^{n} O\,[k]}{\tau}. \tag{2.12}$$

When $O\,[k] = 1$, the sample k was classified as preictal, while when $O\,[k] = 0$ the

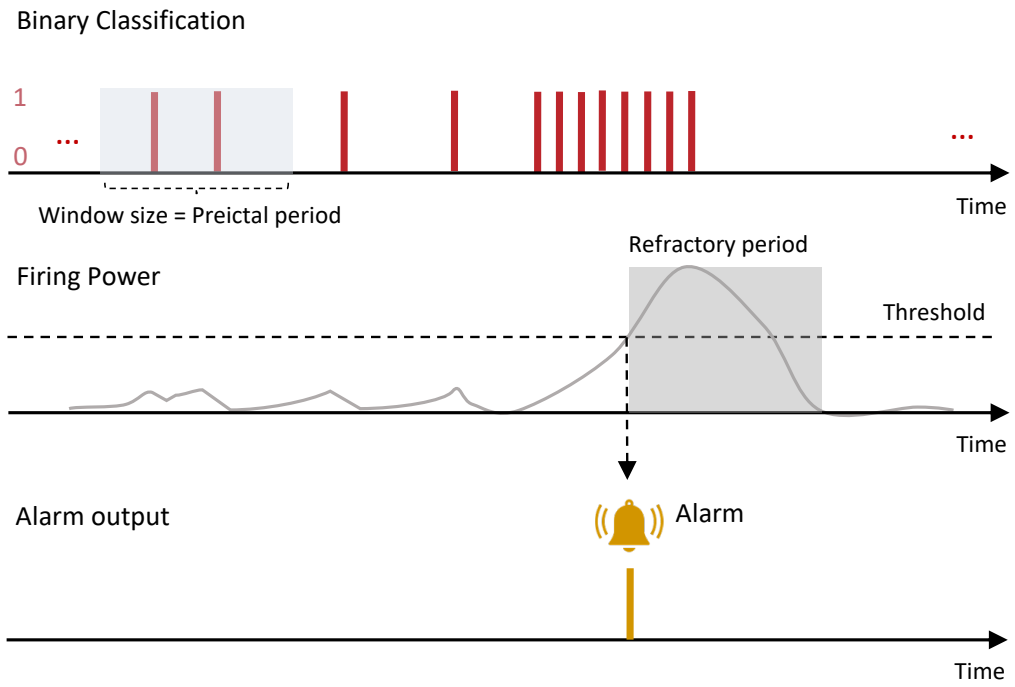


Figure 2.14: Visual representation of the Firing Power. A low-pass filter is applied to the chronological model classifications. An alarm is raised when a certain threshold is passed, followed by a refractory period.

was classified as interictal. $fp[n]$ ranges from zero to one, where an alarm is raised when its value is above a determined threshold (Equation 2.13). The higher the threshold is, the more conservative the alarm generator will be. It is also common to use a refractory period of SOP+SPH duration after each alarm.

$$\begin{cases} alarm\ if\ fp[n] \geq threshold, \\ no\ alarm\ if\ otherwise. \end{cases} \tag{2.13}$$

Although several authors have reported using the Firing Power technique in their studies [Direito et al., 2017, Bandarabadi et al., 2015b, Teixeira et al., 2014b], no optimal threshold was determined. When compared with the Firing Power, the Kalman filter produces more false positives [Assi et al., 2017, Teixeira et al., 2012].

2.5 Concept drifts

An essential aspect of EEG seizure prediction is CDs: the concept of interest may depend on some hidden context that is not given explicitly in the form of predictive features [Lu et al., 2018, Tsymbal et al., 2008]. Additionally, the central assumption with CDs is that the underlying mechanisms that generate new data are unknown to the learner and, thus, unpredictable. In the absence of such knowledge, it is desired that the trained classifier can handle such changes in concepts over time [Hoens et al., 2012].

Changes in the brain dynamics due to daily-life habits, medication, stress situations, and others can induce significant changes in relevant features and data distribution (see Figure 2.15) [Baud et al., 2020, Kuhlmann et al., 2018b, Freestone et al., 2017]. As the majority of EEG databases comprise presurgical monitoring data, patients suffer anti-seizure medication withdrawal and sleep deprivation, which affects the circadian rhythm and the sleep-wake cycles. These changes are visible when comparing the mean seizure frequency between presurgical monitoring and real-life [Winterhalder et al., 2003], or when analysing the tendency to have seizures at periods of sleep transition (going to sleep or waking up) [Kuhlmann et al., 2018b, Freestone et al., 2017].

According to the cycle duration, rhythms can be considered circadian, ultradian, or multidien, as seen in Box 5. Circadian rhythms consist of 24-hour cycles associated with physiological changes, such as the sleep-wake cycle, hormonal production, body temperature, heart rate, and blood pressure. Ultradian cycles include the non-REM-REM, which lasts about 90 minutes. Multidien cycles can be weekly, half-weekly, or last several weeks, where its influence on seizure prediction has been studied as well [Khan et al., 2018, Karoly et al., 2016].

Box 5 - Definition of circadian, multidien and ultradian rhythms (from [Khan et al., 2018]).

Circadian rhythm: *"A biological rhythm is considered to be a circadian rhythm if it meets three criteria: the rhythm should have an endogenous free-running (approximately) 24 h period, should be entrainable (i.e., be capable of phase reset by environmental cues and synchronisation to the 24 h day), and should exhibit temperature compensation."*

Multidien rhythm: *"Refers to rhythms with a time period covering several days."*

Ultradian rhythm: *"Refers to rhythms with periods of less than 24 h; ultradian rhythm cycles can occur with a frequency of more than once per day."*

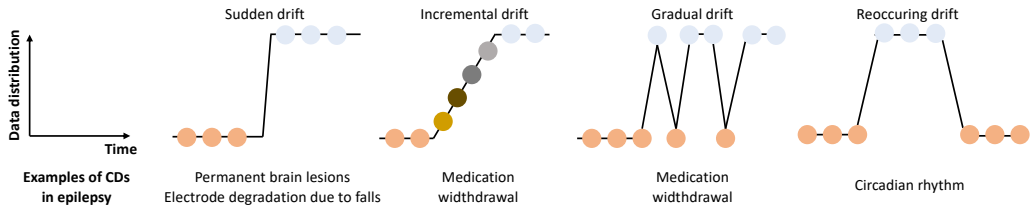


Figure 2.15: Different types of CDs and their possible translation for the specific case of epilepsy seizure prediction, including presurgical monitoring conditions. Adapted from [Lu et al., 2018, Gama et al., 2014].

2.5.1 Sleep-wake cycle

The sleep-wake cycle causes the most significant alteration of behavioural and physiological states, as sleeping occupies about one-third of human life and is essential to preserve physical and mental health. Sleep is a reversible physiological phenomenon that, compared to wakefulness, reduces mobility, conscience, and responsiveness [Tripathy et al., 2020, Chokroverty, 2017].

Sleep is divided into two stages: Non-Rapid Eye Movement (NREM) and Rapid Eye Movement (REM), where the latter cyclically follows the first. Each sleep cycle lasts from 90 to 110 minutes. Adults experience four to six cycles during normal sleep, as in Figure 2.16. The REM stage is characterised, in the EEG, by low-amplitude mixed-frequency signals and beta and theta rhythms. According to the American Academy of Sleep Medicine [Rosenberg and Van Hout, 2013], the NREM phase is divided into three substages: N1 (mainly theta and some delta frequencies and vertex waves), N2 (K-complexes and sleep spindles), and N3 (slow-wave activity, delta rhythms) [Berry et al., 2012, Chokroverty, 2009].

2.6 Explaining models' decisions

With the rise of ML in real-world applications, where healthcare and epilepsy seizure prediction are not exceptions, specific criteria need to be assured due to patient safety requirements. It is necessary to develop high-performing models under rigorous conditions and understand them in a human-comprehensible manner. It is possible to list four main reasons that justify the need for developing a more profound comprehension of the constructed models:

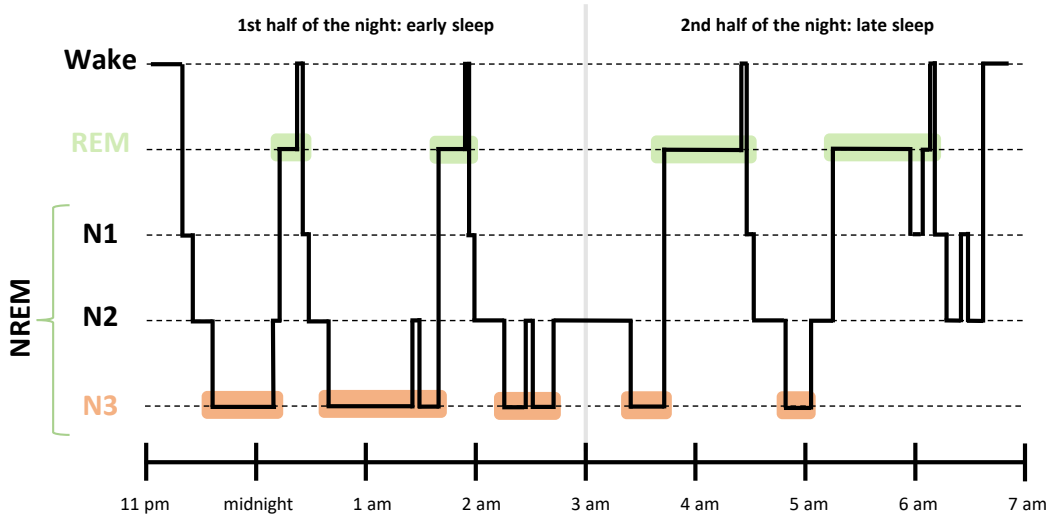


Figure 2.16: An example of a hypnogram that shows the sleep stages in a healthy human within eight hours of sleep. N3 is the longest stage in the early cycles while REM stage increases in each cycle. Adapted from [Blume et al., 2015].

- There is a significant scepticism on ML since these algorithms deal with multi-dimensional inputs and tend to create black-box models. Although clinicians may trust the models when making the right decisions, they may mistrust them when they fail. No model is infallible [Kuhlmann et al., 2018b, Freestone et al., 2017].

- Seizure prediction requires multidisciplinary efforts, where effective communication is necessary between clinicians, researchers, and patients. The clinician must be aware of the models' limitations to make the best decision for the patient. The patient also needs to feel safe as these interventions may increase existing anxiety and fear [Molnar, 2019, Goodman and Flaxman, 2017, Doshi-Velez and Kim, 2017].

- As the EEG is a complex signal, not easy to understand, and optimal pre-seizure biomarkers remain unknown, ML may identify patterns that in long-term data are not perceptible by humans. These need to be explained to clinicians [Chiang et al., 2022, Freestone et al., 2017].

- There is a need for further understanding due to incompleteness in problem formalisation. In specific problems, such as seizure prediction, it is desired not only to predict seizures but also to understand why. A correct prediction may partially solve the original problem, as it is desired to have a more profound knowledge of the brain to help develop better AEDs and improve other neuromodulation techniques [Molnar, 2019, Miller et al., 2017, Doshi-Velez and Kim, 2017], for instance.

In fact, current legislation already demonstrates the need to guarantee the users' safety, as demonstrated by the article 22 (see Box 6) of the General Data Protection Regulation (GDPR), known as *Right to an explanation* [Goodman and Flaxman, 2017]. This article elevates the importance of algorithm explainability for high-risk decisions based on personal data. It emphasises the need for providing patients with the right to have an explanation for any algorithm decision and gives them the right to question those decisions.

Box 6 - The article 22 of the GDPR [Goodman and Flaxman, 2017].

"The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."

Additionally, understanding models' decisions also help to guarantee critical criteria concerning ML use in real-life: i) fairness, the reassurance that there is no discrimination associated with the predictions towards protected groups; ii) robustness, the ability to maintain the same performance levels when submitted to variations of

the input or parameters; and iii) causality, that the model prediction perturbations will also be observed in the real-world system [Doshi-Velez and Kim, 2017].

2.6.1 Interpretability and explainability

In the last years, despite a new scientific area known as interpretability or explainability has emerged, the need for further understanding ML models is not exactly new. It has always been present to some extent. Over the years, many efforts have been made to provide simple models with few intuitive features, particularly sparse linear and logistic regression models and low-depth decision trees [Siddiqui et al., 2020, Mohseni et al., 2006], as these enable accessible communication between researchers and clinicians.

With the advent of *big data* and the exponential increase of computational power over the years, more complex models have more potential to lead to better performances. The area of ML interpretability/explainability arose due to a need for retrieving knowledge of such models [Molnar, 2019, Alkan et al., 2005]. It is important to evidence the distinction between the terms *interpretability* and *explainability* [Gleaves et al., 2020, Gilpin et al., 2018]. Although they may appear synonyms, these have evolved to consider different aspects.

Interpretability concerns a system's ability to show its logic so that a human can predict a future output by only analysing the input. In other words, it offers the simplicity of understanding the model intrinsically so that it is effortless for a human to know the result of a new prediction. Algorithms like k-Nearest Neighbours (kNN), logistic regression, and low-depth decision trees are considered to develop interpretable models. While simple models might lose essential relations, resulting in unsatisfactory performance values, increasing the number of features may increase complexity and consequent loss of interpretability [Molnar, 2019, Gilpin et al., 2018, Kim et al., 2016].

Explainability concerns the ability of a model to explain its reasoning and behaviour in human terms, where it is not necessary to comprehend the underlying mechanisms of the model [Gleaves et al., 2020, Molnar, 2019, Gilpin et al., 2018]. Explainability techniques emphasise understanding a single or set of decisions and not how the model operates intrinsically. Despite one can explain a model's decision by simply providing an interpretable model, it is possible to provide explanations with non-interpretable models. Achieving interpretability is a way to ensure explainability, but not the only one.

2.6.2 What is an explanation

The necessary explanations might differ from problem to problem, where in several cases interpretability is still required [Gilpin et al., 2018]. When exploring explanations for a given problem, the following question is vital: is explainability enough

to ensure applicability in a clinical environment, or is it also required to obtain interpretability? Typically, interpretability is required when medical knowledge is advanced and already performs at a satisfactory level. In those cases, clinicians can more easily ensure patient safety. Explainability is fundamental when dealing with complex underlying knowledge, where humans may not detect such patterns. In these cases, it may be required a model with a complex internal structure [Miller, 2019, Gilpin et al., 2018].

The discussion on the definition of an explanation might be tricky and complex, entering the field of philosophy and linguistics where an explanation can be an exchange of beliefs [Molnar, 2019, Miller, 2019, Miller et al., 2017]. In this thesis, an explanation answers a "why question": why did the model behave in such a way? A good explanation is when one can no longer keep asking why [Molnar, 2019, Gilpin et al., 2018, Miller et al., 2017].

2.6.3 Taxonomy

Methods for explainability can be classified according to several criteria (see Figure 2.17): i) post-hoc or intrinsic, ii) their interpretation results, iii) model-specific or model-agnostic, and iv) local or global [Molnar, 2019].

Intrinsic/post-hoc classification concerns the analysis of developed models. While intrinsic explanations directly analyse the developed models, post-hoc strategies apply other methods to analyse the model after training.

The most common form of classifying explainability methods is according to their interpretation results. According to these criteria, it is likely to find explainability methods grouped in the following form:

- Feature summary statistics and visualisation: depend on the developed clas-

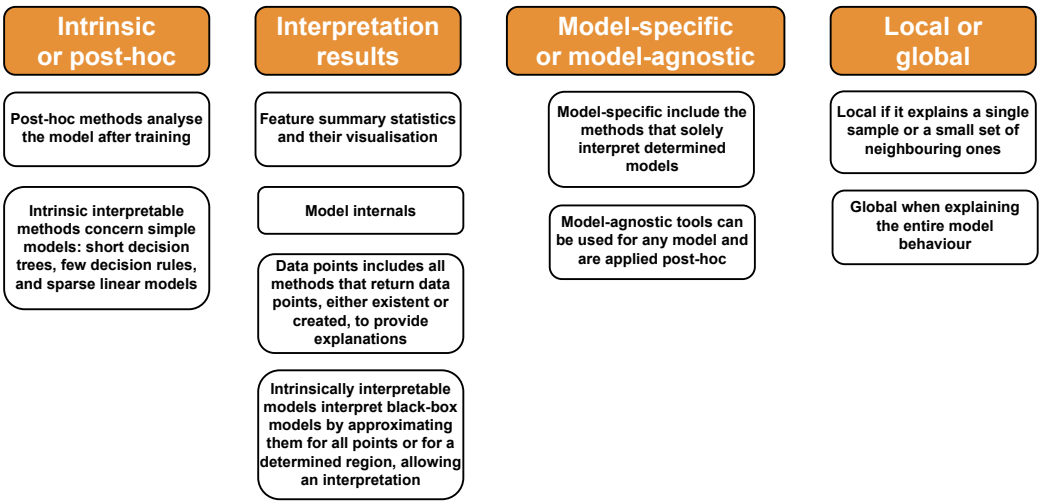


Figure 2.17: The taxonomy of explainability methods. These can be categorised according to different criteria: intrinsic or post-hoc, the interpretation of results, model-specific and agnostic, and their explanation range.

sification model. Examples of summary statistics can be feature importance or feature interaction values [Inglis et al., 2022]. Some of these can also be visualised, where presenting them in a table might not be intuitive. Partial dependence plots [Greenwell et al., 2018] or accumulated local effects [Apley and Zhu, 2020] are curves that show, for a chosen feature, the average outcome [Molnar, 2019]. Other examples are Shapley Values [Štrumbelj and Kononenko, 2014].

- Model internals is also the case of interpreting intrinsically interpretable models. In fact, the distinction between model internals and feature summary statistics might be blurred. Weights in linear models or the learned tree structure are clear examples of it [Molnar, 2019]. Nevertheless, not all cases overlap: model internals are significantly used for interpreting convolutional neural networks, namely the visualisation of feature detectors through saliency maps [Shrikumar et al., 2017, Zeiler and Fergus, 2014, Simonyan et al., 2013].

- Data points: includes all methods that return data points, either existent or created, to provide explanations. One example is counterfactual explanations [Wachter et al., 2017] which find the most similar data points that change the predicted outcome. In other words, a counterfactual explanation describes a causal situation in the form: "if X had not occurred, Y would not have occurred" [Molnar, 2019]. Other cases include identifying relevant points for the obtained outcome, such as influential instances [Koh and Liang, 2017,Cook, 1977]: samples when its deletion from training data considerably changes the predictions from the models, and the detection of important local points using approximate models [Lin et al., 2019].

- Intrinsically interpretable models: it is possible to interpret black-box models by approximating them for all points (globally) or a determined region (locally), using an intrinsically interpretable model. The latter allows an interpretation by inspecting model internals or feature summary statistics. Some examples are Local Interpretable Model-Agnostic Explanations (LIME) [Ribeiro et al., 2016] for text and tabular data.

Explanation methods can also be categorised into model-specific or model-agnostic. Model-specific include the methods that solely interpret determined models, such as saliency maps for neural networks [Shrikumar et al., 2017, Zeiler and Fergus, 2014, Simonyan et al., 2013] or, naturally, interpretations concerning intrinsically interpretable models as these are specific. Model-agnostic tools can be used for any model and are applied post-hoc. As these do not have access to the model logic, they focus on establishing relations between input and output pairs [Molnar, 2019].

Finally, it is also possible to categorise an explanation model according to its range on the dataset. It can be local if it explains a single sample or a small set of

neighbouring ones, or global when explaining the entire model behaviour. In Table 2.3, one can see a list of explainability methods and their categorisation.

2.6.4 Evaluation

Explanations need to be evaluated, which can be performed according to two criteria: interpretability and completeness [Gilpin et al., 2018]. The interpretability goal concerns the tie to the user's cognition, knowledge, and biases: descriptions need to be simple enough and use meaningful vocabulary so that the user understands the provided explanations. Completeness concerns how well the provided explanations accurately describe the system. The more complete an explanation is, the easier it is for the user to predict the system behaviour in more situations.

One of the many challenges in ML explainability is creating explanations that are simultaneously accurate and easy to interpret. Usually, the most interpretable explanations tend to lose predictive power. Evaluating explanations requires particular attention, as human evaluations may imply a strong bias towards simple descriptions. Explanations may lead to systems that tend to be more persuasive than transparent, which may lead to ethical issues by oversimplifying and misleading the user. Authors suggest a tradeoff between interpretability and completeness by making a curve from maximum interpretability to maximum completeness [Gilpin et al., 2018, Herman, 2017].

The evaluation of explanations at the interpretability level can be performed at

Table 2.3: A list of explainability methods, along with its classification according to the different taxonomy criteria.

	Intrinsic or post-hoc		Interpretation results				Model-specific or model-agnostic		Local or global	
	Intrinsic	Post-hoc	Feature summary statistics	Model internals	Data points	Intrinsically Interpretable Models	Specific	Agnostic	Local	Global
Decision Rules	x			x			x			x
Decision tree	x		x	x			x			x
Sparse Linear and Logistic Regressions and Generalised Linear Models	x		x	x			x			x
kNN model	x				x		x			x
Partial Dependence Plot		x	x					x		x
Accumulated Local Effects		x	x					x		x
Feature Interaction		x	x	x			x	x		
Permutation Feature Importance		x	x					x		x
Global Surrogate		x				x		x		x
Prototypes and Criticisms		x			x			x		x
Individual Conditional Expectations		x	x					x	x	
LIME		x			x	x		x	x	
Counterfactual Explanations		x			x			x	x	
Shapley Values		x							x	
SHapley Additive exPlanations		x	x					x	x	
Network Dissection		x		x			x			x
Saliency Maps		x			x		x			x
Adversarial Examples		x			x		x			x
Influential Instances					x			x		x

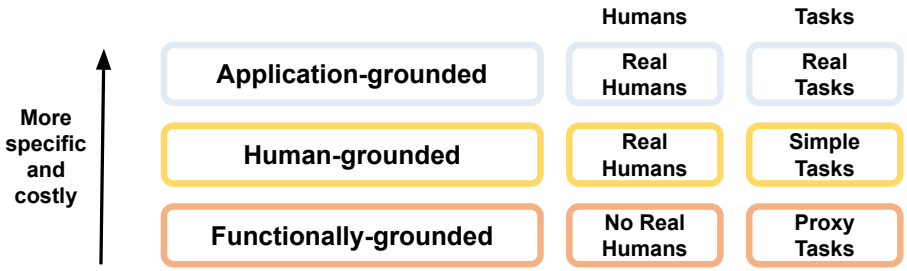


Figure 2.18: The three levels of explanations evaluation concerning interpretability. Adapted from [Doshi-Velez and Kim, 2017].

three levels, as observed in Figure 2.18: application, human, and function [Molnar, 2019, Gilpin et al., 2018, Doshi-Velez and Kim, 2017].

The application-level tests explanations provided to human specialists regarding the actual task. This level aims to understand how efficiently the created explanations help specialists complete their tasks. It the most challenging as it requires several exceptionally trained people. The human-level tests explanations in simplified applications (without jeopardising their core) in humans that are not experts. This level may help represent explanations to patients and, naturally, may also help to handle a possible lack of specialists for application-level evaluations. Lastly, the function level uses a formal and *a priori* definition of interpretability to analyse the explanations without involving humans. This level is the most tested as it does not require humans for testing, and many authors that develop ML methods tend to do it intrinsically when searching for model transparency.

2.7 Summary

Epilepsy is characterised by a significant clinical heterogeneity concerning seizures, type of epilepsy, and epilepsy syndromes. A seizure can be characterised by initial signs/symptoms, awareness, and epileptic focus localisation. The most common epilepsy syndrome is the TLE which is characterised by seizures with temporal lobe focus. DRE patients, which do not achieve sustained seizure freedom through medication, are the focus of seizure prediction as they are exposed to the physical and social implications of the unpredictability of seizures. These patients are often subjected to monitoring for weeks/months to evaluate their condition before undertaking surgical interventions, which explains why most databases comprise data acquired during this period.

The brain's electrical activity can be captured by the EEG which is the primary physician tool, although its morphology is not fully understood. There are two types of acquisitions of EEG: scalp and intracranial. While iEEG has a higher signal-to-noise ratio and captures more trustfully high-frequency band activity, it is an invasive method, presenting a considerable risk of infection/haemorrhage. EEG activity can

be characterised by oscillations or transients. Oscillations are rhythmic patterns, while transients are sharp and can be categorised into normal and abnormal. Within normal transients, they are related to eye blink and muscle impulses, among other normal body functions, while the abnormal can be related to epileptic activity. Not all detected epileptic activity can anticipate seizures. It is necessary to analyse all types of EEG activity to make more accurate predictions.

In seizure prediction, authors divide the EEG signal concerning seizures into interictal, preictal, postictal, and ictal, which is fundamental for supervised learning and evaluating the methodologies' performance. Its goal is to detect the preictal period and anticipate seizures by timely raising alarms. Each alarm is associated with an occurrence period (SOP) and an intervention time (SPH). However, as the preictal period represents a transitional stage that varies within patients and seizure episodes, it can be challenging to detect this interval.

A proper evaluation for an alarm system should assess seizure sensitivity and FPR/h. A proper methodology must have an adequate SPH and SOP to allow a given intervention in real-life. Statistical validation should also be conducted where performing above chance is a minimum requirement. As in any other rare-event prediction task within a time series, proposed approaches must deal with data imbalance and concept drifts. The most common presurgical monitoring concept drifts are the circadian cycle, sleep-wake cycle, and medication tapering.

Explaining prediction models' decisions is a fundamental task to allow clinical acceptance by reassuring patient safety and dealing with ML scepticism. Furthermore, as the EEG is a complex signal, there is the possibility to show clinicians patterns not previously identified as seizure predictive. It is also desired to have a more profound knowledge of the brain to help develop better AEDs and improve other neuromodulation techniques. A good explanation answers a "why" question.

Chapter 3

State of the Art

This chapter provides an overview of the state of the art in seizure prediction over the past ten years, mainly based on the Electroencephalogram (EEG) signal. Section 3.1 presents the most common framework. Section 3.2 describes the application of concept drifts in prediction. Section 3.3 presents an overview in explainability of EEG-based models. Lastly, section 3.4 summarises the state-of-the-art key concepts and provides final reflections.

3.1 Seizure Prediction

Current seizure prediction algorithms have a common framework that consists of signal acquisition, preprocessing, feature extraction, feature selection, classification, regularisation, and performance evaluation, as schemed in Figure 3.1. These steps can be summarised as follows:

- Signal preprocessing enhances the EEG quality and extracts signal information through sliding time-window analysis.
- Feature extraction and selection collect characteristics that are expected to be sensitive to the detection of pre-seizure generation mechanisms.
- Classification consists of training Machine Learning (ML) models with the previously selected features to identify periods as either interictal or preictal.
- Regularisation smooths the classification output by not considering isolated classifications and provides them temporal meaning.

Despite a common framework, the variety of possibilities is significantly large, explaining the heterogeneity of existing approaches. The absence of a gold standard algorithm also contributes to this.

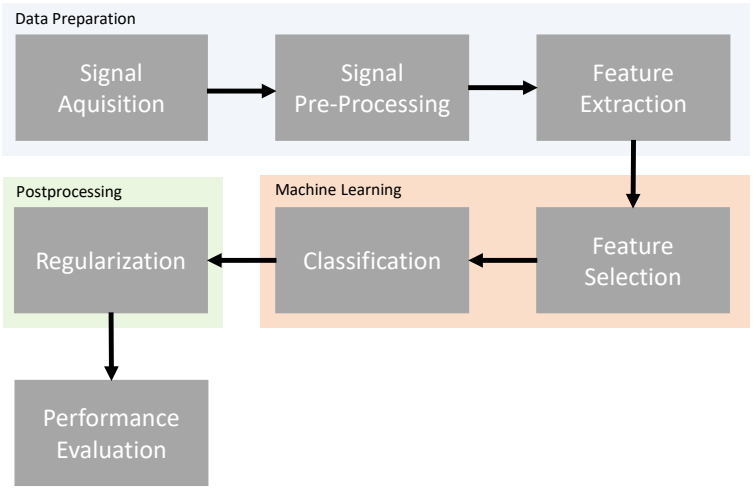


Figure 3.1: Common framework of prediction algorithms. Adapted from [Assi et al., 2017].

Differences in Deep Learning approaches

With the increase of computational power and available data over the years, more advanced ML models can be explored, as in the case of Deep Learning (DL) models. These models also can handle raw data by automatically performing preprocessing and feature engineering, able to enhance classification [Kuhlmann et al., 2018b, Freestone et al., 2017]. Due to this, several modifications can be made to the common framework, as seen in Figure 3.2 where three alternatives are presented.

The most straightforward alternative (A in Figure 3.2) is to provide the raw input (which can have a low processing level) to a DL model and retrieve the classification output throughout the sequence analysis. In this case, the model is responsible for signal processing (artefact removal, noise reduction, and filtering), feature engineering, and classification [Viana et al., 2022, Pal Attia et al., 2022, Xu et al., 2020, Zhang et al., 2019, Truong et al., 2018, Tsiouris et al., 2018].

Instead of using raw data as input, some authors previously perform feature extraction and provide the obtained measures to the models (B in Figure 3.2). In this case, the model is expected to focus on dimensionality reduction, feature selection, and classification [Stirling et al., 2021b]. Lastly, some authors use these models as feature engineering by extracting the obtained coefficients, which are provided to another classifier (C in Figure 3.2) [Usman et al., 2021b, Daoud and Bayoumi, 2019].

3.1.1 Signal acquisition

Table 3.1 shows an overview of the used data in seizure prediction studies over the last ten years. Although this analysis mainly concerns the EEG signal, other signals are included, such as blood volume pulse, accelerometry, electrodermal activity, and sleep [Stirling et al., 2021b, Nasseri et al., 2021]. These studies highlight a

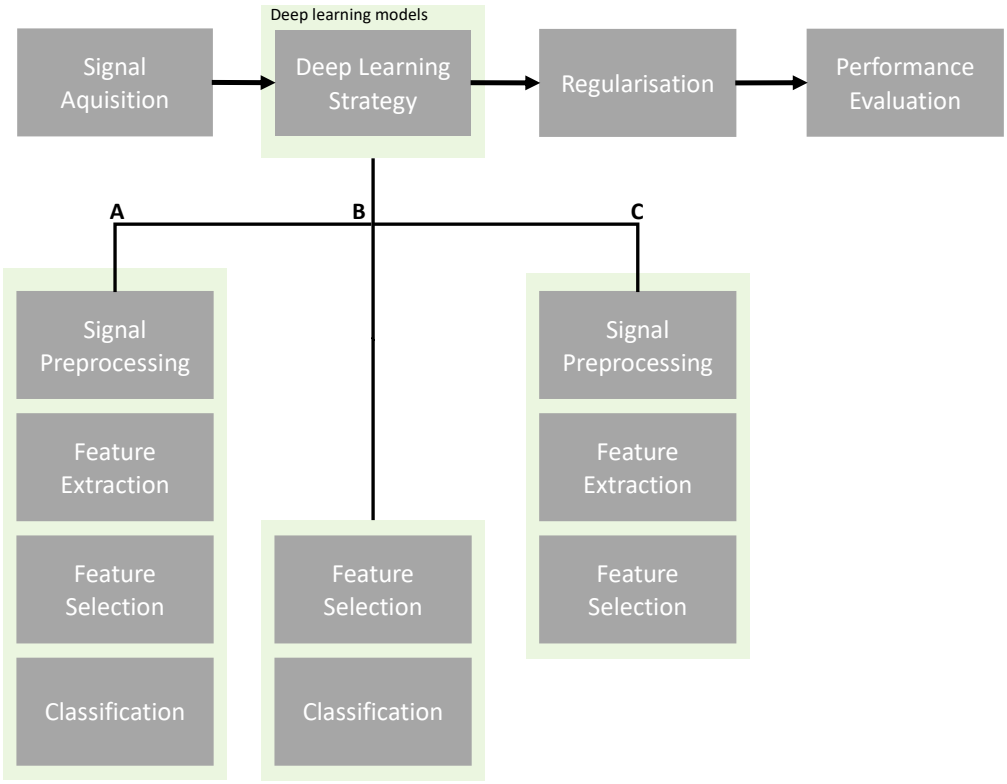


Figure 3.2: Possible variations to the common framework of prediction algorithms when using DL approaches. Green represents the use of DL models. Adapted from [Assi et al., 2017].

possible paradigm change resulting from the emergence of seizure forecasting and the possibility of acquiring recordings during more extended periods (several months or even years per patient), which require more comfortable strategies to acquire physiological information.

In an early stage, studies were mainly based on local databases acquired from patients undergoing evaluation for epilepsy surgery, as the case of the EPILEPSIAE, Freiburg, and CHB-MIT databases. These are still used.

[Cook et al., 2013] published the first study using the Neurovista database, containing chronic Invasive Electroencephalogram (iEEG) (up to two years per patient). This data was then used in several studies, such as [Kiral-Kornek et al., 2018], where the three worst-performing patients were used in a Kaggle Competition [Kuhlmann et al., 2018a]. Due to comfort issues, other ultra-long-term databases are arising through wrist-worn bands and smartwatches. These studies are more likely to be successful: they use more data, study concept drifts, and provide higher patient comfort. Additionally, other EEG databases are arising, such as the one from SeizeIT2 [Zhang et al., 2022] (NCT04284072) clinical trial, and using subscalp EEG, such as the clinical trial from King's College London (NCT04061707), and the Zealand University Hospital [Weisdorf et al., 2019]. Other databases are related to another Kaggle contest comprised of dogs and humans [Xu et al., 2020, Truong

Table 3.1: Overview of the signal acquisition from seizure prediction over the past ten years.

Study	Database	Patients (Aggregated analysed time) No. of Seizures	Signal	Electrodes
[Viana et al., 2022]	ZUH KCL's clinical trial	6 (594 days) 82	Subcutaneous EEG	-
[Pal Attia et al., 2022]	ZUH KCL's clinical trial	6 (409 days) N.A.	Subcutaneous EEG	-
[Usman et al., 2021b]	CHB-MIT	23 (27 days) 198	Scalp EEG	-
[Stirling et al., 2021b]	Personal	11 (13.5 years) 1493	BVP Sleep stages	Smartwatch
[Nasseri et al., 2021]	NeuroPace	6 (4 years) 278	ACC BVP, EDA TEMP	Wrist-worn band
[Xu et al., 2020]	Kaggle iEEG dataset CHB-MIT	5 dogs + 22 (1.85 + 1.18 days) 44+45	iEEG Scalp EEG	-
[Zhang et al., 2019]	CHB-MIT	22 (-) 182	Scalp EEG	-
[Truong et al., 2019]	Freiburg CHB-MIT EPILEPSIAE	13+13+30 (12.95+8.7+120 days) 59+64+261	Scalp EEG iEEG	6, 22, 19
[Daoud and Bayoumi, 2019]	CHB-MIT	8 (-) 43	Scalp EEG	-
[Kiral-Kornek et al., 2018]	NeuroVista	15 (16.29 years) 2817	iEEG	16
[Tsiouris et al., 2018]	CHB-MIT	12 (40 days) 185	Scalp EEG	-
[Truong et al., 2018]	Freiburg CHB-MIT Kaggle (AES)	28 + 2 dogs + 6 (13+8.7+26 days) 59+64+48	Scalp EEG iEEG	6, 22, 19
[Kuhlmann et al., 2018a]	Neurovista	3 (442 days) 211	iEEG	16
[Karoly et al., 2017]	Neurovista	9 (10.35 years) 1458	iEEG	-
[Direito et al., 2017]	EPILEPSIAE	216 (697t days) 1206t	Scalp EEG iEEG	F7, FZ, F8, T5, PZ, T6 6 random 6 in focal region
[Bandarabadi et al., 2015b]	EPILEPSIAE	24 (150t days) 183t	Scalp EEG iEEG	3 in focal region and 3 far from local region
[Assi et al., 2015]	Kaggle (AES)	5 dogs (-) 44	iEEG	16
[Rasekhi et al., 2015]	EPILEPSIAE	10 (58 days) 86	Scalp EEG iEEG	3 in focal region and 3 far from local region
[Teixeira et al., 2014b]	EPILEPSIAE	278 (2031 days) 2702	Scalp EEG iEEG	F7, FZ, F8, T5, PZ, T6 6 random 6 in focal region
[Alvarado-Rojas et al., 2014]	EPILEPSIAE	53 (531 days) 558	iEEG	-
[Moghim and Corne, 2014]	Freiburg	21 (24 days) -	iEEG	3 in focal region and 3 far from local region
[Rasekhi et al., 2013]	EPILEPSIAE	10 (31t days) 46t	Scalp EEG iEEG	3 in focal region and 3 far from local region
[Rabbi et al., 2013]	EPILEPSIAE	1 (1.5 days) 7	iEEG	2
[Cook et al., 2013]	Neurovista	15 (≈ 16 years) 1392	iEEG	16

AES stands for American Epilepsy Society. CHB-MIT for the Children's Hospital Boston from the Massachusetts Institute of Technology, ZUH for Zealand University Hospital, and KCL for King's College London. In analysed time and seizures, "t" stands for testing data. BVP, ACC, EDA, and TEMP stand for blood volume pulse, accelerometry, electrodermal activity, and temperature.

et al., 2018].

Electrode selection

Despite the majority of scalp EEG databases being acquired within the 10-20 system, some authors do not use all electrodes to simulate a real-life application and to increase patient comfort. While some use all available electrodes [Usman et al., 2021b, Zhang et al., 2019, Daoud and Bayoumi, 2019, Tsiouris et al., 2018], others choose six where three belong to the focal region and the remaining are placed far from it [Bandarabadi et al., 2015b, Rasekhi et al., 2015, Rasekhi et al., 2013]. Others focused only on electrodes from the focal region [Direito et al., 2017, Teixeira et al., 2014b]. These choices lead to different assumptions that are worth investigating.

For instance, by choosing random electrodes, one assumes the seizure generation processes can be captured in any brain location. By choosing three electrodes near the focal region and three far from it, the authors assume that it is necessary to relate information from the focus lobe to other brain regions without using all possible electrodes. When only electrodes from focal regions are used, the assumption is that the activity of the focal region is enough to capture a seizure-generation process. No assumption has proved to be the best, while it is intuitive that using all electrodes available may provide more information.

3.1.2 Preprocessing

Signal preprocessing aims to enhance the EEG quality and extract information. As the main objective is to construct a method to receive and process online data, the chosen methods must consider their real-life feasibility. The first step comprises data segmentation by window analysis. Then, some other options can be performed as denoising, filtering, artefact removal, and decomposition [Direito et al., 2017, Bandarabadi et al., 2015b, Assi et al., 2015, Rasekhi et al., 2015, Teixeira et al., 2014b, Rabbi et al., 2013, Rasekhi et al., 2013, Park et al., 2011, Chisci et al., 2010].

Then, one must define the preictal period, Seizure Occurrence Period (SOP), and Seizure Prediction Horizon (SPH). These steps are not necessarily included in this stage as they can be handled in the classification tasks. Nevertheless, they are presented here as they should be considered at the beginning of the study, so that this choice does not influence the performance of the ML model. Figure 3.3 presents a general pipeline for the signal preprocessing stage.

Table 3.2 provides a general overview of the authors' decisions concerning signal preprocessing. Summarily, denoising, filtering, and artefact removal are steps where most studies do not concentrate significant efforts, as the EEG is a complex signal and hard to understand. Using these intense preprocessing strategies carries the risk of losing relevant brain information.

Table 3.2: Overview of the signal preprocessing steps, preictal period, and SPH duration over the last ten years.

Study	Sliding window	Filtering	Preictal Period	SPH
[Viana et al., 2022]	1 min No overlap	0.5-48Hz band-pass and 25Hz low-pass filters 40dB attenuation filter	1 hour	-
[Pal Attia et al., 2022]	1 min No overlap	0.5-48Hz band-pass filter 40dB attenuation filter	1 hour	5 min
[Usman et al., 2021b]	29s No overlap	Empirical Mode Decomposition	32 min	-
[Stirling et al., 2021b]	5s and 1 min No overlap	Butterworth band-pass filter Hilbert transform	1 hour and 24 hours	-
[Nasseri et al., 2021]	1s and 4s	-	1 hour	15 min
[Xu et al., 2020]	20s No overlap	-	30 min	5 min
[Zhang et al., 2019]	5s No overlap	5th-order Butterworth band-pass filter 5-50Hz	30 min	-
[Truong et al., 2019]	28s No overlap	Band-pass filters as notch filters 47-53Hz and 97-103Hz	30 min	5 min
[Daoud and Bayoumi, 2019]	5s No overlap	-	60 min	-
[Kiral-Kornek et al., 2018]	5s No overlap	Octave-wide digital and notch filters 8Hz-128Hz	15 min	-
[Tsiouris et al., 2018]	5s No overlap	-	15, 30, 60, 120 min	-
[Truong et al., 2018]	30s No overlap	Notch-Filters DC removed	30 min	5 min
[Kuhlmann et al., 2018a]	0s to 10min 0 to 50% overlap	-	55 min	5 min
[Karoly et al., 2017]	60s 50% overlap	1-140Hz band-pass filter	30 min	1 min
[Direito et al., 2017]	5s No overlap	50Hz notch filter	10:10:40 min	10s
[Bandarabadi et al., 2015b]	5s No overlap	50Hz notch filter	10:10:40 min	-
[Assi et al., 2015]	5s No overlap	50Hz notch 0.5 - 180Hz band-pass	60 min	5s
[Rasekhi et al., 2015]	5s No overlap	50Hz notch filter	10:10:40 min	-
[Teixeira et al., 2014b]	5s No overlap	50Hz notch filter	10:10:40 min	10s
[Alvarado-Rojas et al., 2014]	5s No overlap	8th-order Butterworth filter in bands of interest from 0.5Hz to 140Hz Hilbert transform	60 min	1 min
[Moghim and Corne, 2014]	5s and 9s No overlap	Artefact removal with EEGLAB	5 min	-
[Rasekhi et al., 2013]	5s No overlap	50Hz notch filter	10:10:40 min	-
[Rabbi et al., 2013]	10 seconds 50% overlap	60Hz notch 0.5 - 100Hz band-pass	15, 30, 45 min	-
[Cook et al., 2013]	5s No overlap	Octave-wide digital and notch filters 8Hz-128Hz	minutes to hours	-

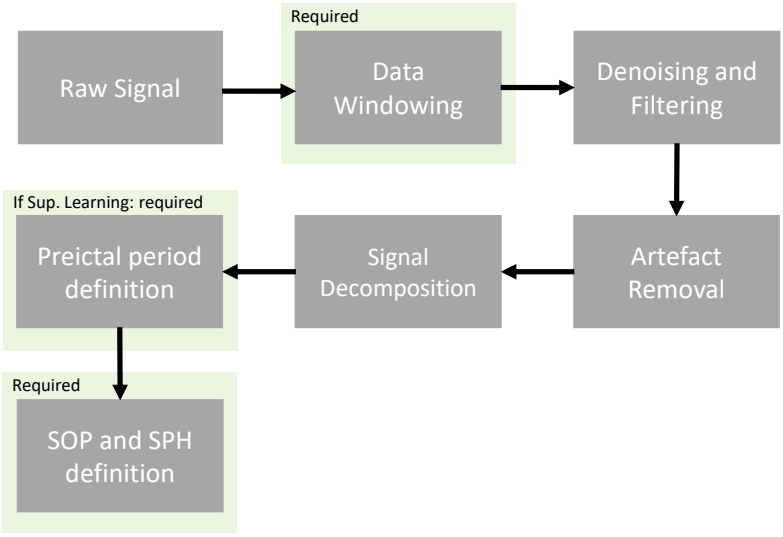


Figure 3.3: Flowchart of a common signal preprocessing pipeline. Required steps are in green. Defining the preictal period is required when using a supervised learning approach.

Data segmentation

The EEG signal must be segmented into small windows to extract features chronologically to simulate an online time-series analysis. This window length has varied in the literature, generally ranging from 2 to 60 seconds, where five seconds is the most adopted interval. This interval has as objective a contextual meaning according to EEG clinical characteristics.

Due to the number of electrodes, sampling frequency, and the recording duration, the choice of window length and overlap percentage considers computational efforts and execution speed, along with a compromise between capturing specific patterns and stationarity assumptions [Bandarabadi et al., 2015b, Teixeira et al., 2014b, Rasekhi et al., 2013, Park et al., 2011, Chisci et al., 2010].

Denoising, filtering, and artefact removal

This step generally removes the powerline interference, bandpass filtering, and abnormal transients, considered artefacts. Frequency decomposing into the frequency bands of interest or wavelet decomposition can also be considered filtering and artefact removal methods, as other activity patterns besides EEG oscillations may be removed [Moghim and Corne, 2014, Park et al., 2011, Adeli et al., 2007].

Authors have differed in the bandpass filter cut-off frequencies. Generally, they remove low-frequency components below 0.5 Hz, considered breathing artefacts, and high-frequency components, considered noise. The limit for high frequencies varies since many researchers have found discriminative ability with these [Zhang et al., 2019, Kiral-Kornek et al., 2018, Assi et al., 2017, Bandarabadi et al., 2015b, Rabbi et al., 2013, Cook et al., 2013].

Although not explicit, many authors do not put much effort into this stage due to the underlying assumption that signal decomposition, feature extraction, and data model adaptation may be robust to noise and artefacts in the classification stage. Signal decomposition can also be considered part of the feature extraction process.

Preictal period duration, SOP, and SPH

An optimal preictal duration has not been found so far. Authors have adopted fixed periods comprising 2, 20, 30, 60, or even 240 minutes. These can either be previously fixed for all patients [Viana et al., 2022, Pal Attia et al., 2022, Nasseri et al., 2020, Zhang et al., 2019, Kiral-Kornek et al., 2018, Truong et al., 2019, Kuhlmann et al., 2018a, Karoly et al., 2017, Alvarado-Rojas et al., 2014] or defined with a grid-search procedure [Direito et al., 2017, Rasekhi et al., 2015, Teixeira et al., 2014b, Rasekhi et al., 2013, Rabbi et al., 2013]. Other possible solutions can be further applied to prediction, such as unsupervised learning to determine the preictal labels [Müller et al., 2022, Leal et al., 2021].

SPH duration is often omitted in studies [Usman et al., 2021b, Xu et al., 2020, Daoud and Bayoumi, 2019, Kiral-Kornek et al., 2018, Tsiouris et al., 2018, Bandarabadi et al., 2015b, Rasekhi et al., 2015, Rasekhi et al., 2013]. The latter is a significant limitation of the correspondent studies. It represents an unrealistic scenario in real-life applicability, as it is unclear if there is a time to render an intervention.

3.1.3 Feature extraction

Feature extraction is the most heterogeneous step, where no specific type of features has been determined as optimal. These usually aim at capturing one of the following behaviours that concern a change in a pre-seizure state [Assi et al., 2017, Rasekhi et al., 2013, Mormann et al., 2007]: i) an increase in energy caused by electrical discharges in the brain; ii) a shift in spectral power from lower to higher frequencies; and iii) an increase of neuronal synchronisation.

This step can be performed on a single (univariate) or multi-channel basis (bivariate or multivariate). The single-channel analysis selects a given electrode and is mainly based on local activity measures. Multi-channel provides more information as it incorporates the information from two or more electrodes [Assi et al., 2017, Gadhoumi et al., 2016b].

Features can also be classified into linear or nonlinear, where several studies addressed their differences in performance. No conclusion can be drawn as results were conflicting [Harrison et al., 2005, Mormann et al., 2005, McSharry et al., 2003]. Additionally, nonlinear features may not be suited to online processing as their computational cost is higher [Assi et al., 2017, Park et al., 2011].

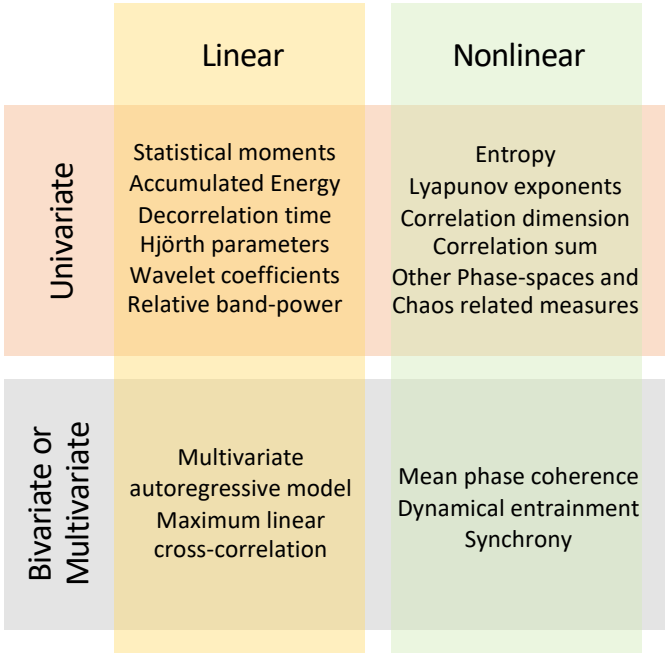


Figure 3.4: A common feature extraction categorisation in terms of linearity and uni/bi/-multivariate analysis into linear univariate, nonlinear univariate, linear bi/multivariate, and nonlinear bi/multivariate features, with some examples.

Grouping features according to linear/nonlinear and univariate/bivariate/multivariate is often seen to organise feature extraction, as presented in Figure 3.4.

Table 3.3 presents a feature extraction overview of linear and nonlinear univariate and multivariate measures. The set of chosen studies shows that linear univariate features are more common than nonlinear univariate ones. These measures are preferred because they are computationally lighter and easier to interpret. Multivariate measures generally require more computational power.

Concerning DL models, some authors use them to perform feature engineering that may be independent of classification. It is not uncommon to see researchers using Convolutional Neural Networks (CNN) to extract features, followed by classification using Long Short-Term Memory (LSTM) networks as these handle time directly [Usman et al., 2021b, Daoud and Bayoumi, 2019]. When the goal is to understand more about the underlying problem, it may be preferable to extract hand-crafted features compared to automatically extracted ones.

The most adopted features can be categorised into linear univariate, nonlinear univariate, linear bi/multivariate, and nonlinear bi/multivariate. A description of these features can be found in Appendix B

Table 3.3: Overview of the used features from seizure prediction over the past ten years.

		Linear univariate features							Nonlinear univariate features						Linear bi/multivariate features		Nonlinear bi/multivariate features		
Study	**Other**	**Statistical moments**	**Spectral band related**	**Wavelets**	**Linear modelling**	**Energy**	**Hjorth parameters**	**Decorrelation time**	**Phase-space and Chaos**	**Lyapunov exponent**	**Dynamic Similarity Index**	**Line-length**	**Energy**	**Entropy**	**Ratio**	**Correlation**	**Dynamical entrainment**	**Mean Phase Coherence**	**Synchrony**
[Viana et al., 2022]	Raw data and FFT data		X																
[Pal Attia et al., 2022]	Raw data, FFT data, and TOD		X																
[Usman et al., 2021b]	From raw data to STFT		X																
[Stirling et al., 2021b]	HR features Time of the day Sleep features																		
[Nasseri et al., 2021]	Raw data HR Time of the day																		
[Xu et al., 2020]	Raw data																		
[Zhang et al., 2019]	From raw data to CSP																		
[Truong et al., 2019]	From raw data to STFT		X																
[Daoud and Bayoumi, 2019]	Raw data																		
[Kiral-Kornek et al., 2018]	From raw data to Spectograms Time of the day		X																
[Tsiouris et al., 2018]	Raw data																		
[Truong et al., 2018]	From raw data to STFT		X																
[Kuhlmann et al., 2018a]		X	X	X	X	X	X	x	X				X					X	X
[Karoly et al., 2017]						X						X	X						
[Direito et al., 2017]		X	X	X	X	X	X	X											
[Bandarabadi et al., 2015b]			X												X				
[Assi et al., 2015]			X				X	x											
[Rasekhi et al., 2015]		X	X	X	X	X	X	x							X				
[Teixeira et al., 2014b]		X	X	X	X	X	X	x											
[Alvarado-Rojas et al., 2014]	Phase interaction with HFO																		
[Moghim and Corne, 2014]		X	X	X		X			X	X									
[Rasekhi et al., 2013]		X	X	X	X	X	X	x											
[Rabbi et al., 2013]											X							X	x
[Cook et al., 2013]						X						X	X						

STF stands for Short-Time Fourier Transform, FFT for Fast-Fourier Transform, TOD for Time Of the Day, HR for Heart Rate, CSP for Common Spatial Patterns, and HFO for High-Frequency Oscillations.

3.1.4 Feature selection

Extracting many features creates a high-dimensional space that needs to be reduced due to the cost of training with the entire feature dataset. Additionally, using a small number of features may improve classification performance as it avoids overfitting. Some features may be redundant, while others can even confound and degrade classifier performance as their discriminative power is not so effective as others.

The rationale behind any feature selection method is to maximise relevance by selecting features with higher discriminative power while minimising similarity by deleting redundant features. ReliefF [Moghim and Corne, 2014], minimum normalized difference of percentiles, maximum Difference Amplitude Distribution histograms (mDAD) [Bandarabadi et al., 2015a] and minimum Redundance Maximum Relevance (mRMR) [Assi et al., 2017, Rasekhi et al., 2015, Bandarabadi et al., 2012] are some methods adopted in prediction studies.

Genetic Algorithms (GAs) are also used and tend to replicate the principles of biological evolution: start from an initial and random population where the strongest will combine to survive and adapt to their external environment. Several strategies have been used, varying in the selection method, genetic structure, and fitness function [Assi et al., 2015, Direito et al., 2011, D'Alessandro et al., 2003].

Furthermore, one can adopt an alternative strategy through dimensionality reduction, such as Principal Component Analysis (PCA) [Acharya et al., 2013, Assi et al., 2017], able to transform high-dimensional data into a low-dimensional orthogonal feature subspace. Each of the orthogonal features is a principal component. The eigenvalues of the data covariance matrix then provide the order of the principal components.

In DL approaches, reduction is performed either by convolutional layers [Usman et al., 2021b, Xu et al., 2020, Truong et al., 2019] or through autoencoders [Daoud and Bayoumi, 2019].

3.1.5 Classification

A prediction model is trained to distinguish preictal and interictal samples based on the extracted features. Authors have explored several models, where a transition from Support Vector Machiness (SVMs) [Assi et al., 2017] to CNNs [Kuhlmann et al., 2018b, Freestone et al., 2017] and LSTMs [Viana et al., 2022, Pal Attia et al., 2022, Usman et al., 2021b, Nasseri et al., 2021, Daoud and Bayoumi, 2019, Tsiouris et al., 2018] has been observed. Other classifiers have been employed, such as random forests and decision trees [Stirling et al., 2021b, Kuhlmann et al., 2018a, Cook et al., 2013], k-Nearest Neighbourss (kNNs) [Cook et al., 2013], and generalised linear models [Stirling et al., 2021b, Karoly et al., 2017]. In Table 3.4, one can see an overview of classification and performance evaluation steps.

An existing problem in training a model is the data imbalance, as interictal sam-

Table 3.4: Overview of the classification, regularisation, performance, and statistical validation over the past ten years.

Study	Training data (testing data)	Classification (regularisation)	Performance	Statistical Validation
[Viana et al., 2022]	Initial 1/3 of data (last 2/3 of data)	LSTM 1h smooth	SS=0.73 TiW=0.34	5 in 6 (0.83) Surrogate analysis
[Pal Attia et al., 2022]	k-fold cross validation with patients	LSTM 1h smooth	SS=0.54 TiW=0.33	4 in 6 (0.67) Surrogate analysis
[Usman et al., 2021b]	k-fold cross validation with seizures	CNN+LSTM	SS=0.93 SP=0.92	No
[Stirling et al., 2021b]	Retraining and testing chronologically and iteratively	LSTM+Random Forest+Log Reg (Kalman Filter)	AUC=0.74	11 in 11 (1.00) Random Forecast
[Nasseri et al., 2021]	First 2/3 of data (last 1/3 of data)	LSTM (Kalman Filter)	AUC=0.80	5 in 6 (0.83) Random Predictor
[Xu et al., 2020]	80% samples (20% samples)	CNN	SS=0.96 FPR/h=0.07	No
[Truong et al., 2019]	Leave-One-Out with seizures	GAN, CNN, NN	AUC=0.81	51 in 56 (0.91) Hanley-McNeil AUC test
[Zhang et al., 2019]	Leave-One-Out with seizures	CNN (Kalman Filter)	SS=0.92 FPR/h=0.12	Statistical comparison between methods
[Daoud and Bayoumi, 2019]	Leave-One-Out with seizures	CNN, Bi-LSTM	SS=0.99 FPR/h=0.004	No
[Kiral-Kornek et al., 2018]	First 2 months (remaining duration)	CNN	SS=0.69 FPR/h=0.00	15 in 15 (1.00) Random Predictor
[Tsiouris et al., 2018]	K-fold with recordings	LSTM	SS=0.99 FPR/h=0.02	No
[Truong et al., 2018]	Leave-One-Out with seizures	CNN (Kalman Filter)	SS=0.79 FPR/h=0.14	28 in 31 (0.90) Random Predictor
[Kuhlmann et al., 2018a]	Training and testing clips	GLMs, SVM, CNN Ensembles, Boosting, Trees	AUC=0.75 FPR/h=0.58	No
[Karoly et al., 2017]	Day 100-200 (Day 200 onwards)	Logistic Regression (Bin width of 1h)	SS=0.60 TiW=0.23	9 in 9 (1.00) Time-matched predictor
[Direito et al., 2017]	2 - 3 seizures / patient (Remaining seizures)	SVM (Firing Power)	SS=0.38 FPR/h=0.20	24 in 216 (0.11) Random Predictor
[Bandarabadi et al., 2015b]	First 3 seizures / patient (Remaining seizures)	SVM (Firing Power)	SS=0.76 FPR/h=0.10	23 in 24 (0.96) Random Predictor
[Assi et al., 2015]	80% segments (Remaining segments)	SVM, ANFIS	SS=0.85 SP=0.80	No
[Rasekhi et al., 2015]	First 3 seizures / patient (Remaining seizures)	SVM (Firing Power)	SS=0.61 FPR/h=0.11	5 in 10 (0.50) Random Predictor
[Teixeira et al., 2014b]	2 - 3 seizures / patient (Remaining seizures)	SVM, ANN (Firing Power)	SS=0.74 FPR/h=0.28	Statistical comparison between methods
[Alvarado-Rojas et al., 2014]	First 4 seizures / patient and at least 10 hours of data (Remaining seizures)	Thresholding (Kalman Filter)	SS=0.68 FPR/h=0.33	7 in 53 (0.13) Random Predictor
[Moghim and Corne, 2014]	10-fold cross validation with 70%/30% samples	SVM	SS=0.91 SP=1.00	Unspecific predictors
[Rasekhi et al., 2013]	First 3 seizures / patient (Remaining seizures)	SVM (Firing Power)	SS=0.74 FPR/h=0.15	No
[Rabbi et al., 2013]	1 seizure / patient (5 seizures)	ANFIS	SS=0.80 FPR/h=0.46	No
[Cook et al., 2013]	First 4 months (Remaining duration)	kNN+Decision Tree (Smoothing)	SS=0.61 TiW=0.23	9 in 10 (0.90) Time-matched predictor

ples are significantly more abundant than preictal ones. Authors have tackled this issue either by: i) balancing the data [Bandarabadi et al., 2015b, Teixeira et al., 2014b, Rasekhi et al., 2013, Direito et al., 2017] by discarding some of the interictal samples; ii) adapting the classifiers and transforming them into cost-sensitive to handle these imbalances [Assi et al., 2017, Park et al., 2011], or iii) artificially generating new preictal samples through strategies such as generative adversarial networks (GANs) [Truong et al., 2019].

Data partition strategies

Several data partition methods have been adopted. Authors should not use segments from the same ictal-related event for training and testing as a clear bias may exist. For instance, by choosing random samples for training and testing in a time series, one is very likely to obtain high performance. This performance overestimation can happen since the trained model learned with neighbouring samples from the tested

ones.

The chosen partition method necessarily leads to different assumptions. Some authors [Acharya et al., 2013, Bandarabadi et al., 2012] choose a determined number of seizures from all patients for training while the remaining are used in the test phase. Others may also choose some patients for training and the remaining for testing. With these strategies, researchers assume that seizure generation processes are not patient-specific. This strategy has been abandoned over time due to unsatisfactory results and due to the great patient heterogeneity [Kuhlmann et al., 2018b, Freestone et al., 2017]. Most studies used patient-tailored methods (see Table 3.4) where a model is trained and tested within each patient data [Direito et al., 2017, Moghim and Corne, 2014, Rabbi et al., 2013].

Authors assume the necessity to use the first seizures for training and the remaining for testing [Teixeira et al., 2014b, Alvarado-Rojas et al., 2014, Teixeira et al., 2012]. Studies performed with ultra-long-term recordings (lasting at least months for each patient) assume concept drifts, where authors deal with them directly by periodically retraining their classifiers [Nasseri et al., 2021, Kiral-Kornek et al., 2018, Cook et al., 2013].

SVMs

SVMs can produce nonlinear decision boundaries and are known to have good generalisation capabilities, as they can transform the feature space into a higher-order one to linearise the data [Assi et al., 2017, Rasekhi et al., 2015, Teixeira et al., 2014b, Rasekhi et al., 2013].

The linearisation is implicitly done using kernel functions and vector operations where the Gaussian Radial Basis Function (RBF) is one of the most used kernels for handling nonlinear problems. By considering σ the scale parameter that is related to the Gaussian width, and x and x' two different feature vectors in the original input, the RBF can be defined as in Equation 3.1:

$$K(x, x') = exp(\frac{-\|x - x'\|^2}{2\sigma^2}). \tag{3.1}$$

This operation replaces each point in the feature space by the gaussian of the squared euclidean distance from support vectors. Two scale parameters need tuning: σ and the soft margin C. The latter controls the trade-off between margin-width maximisation and misclassified samples minimisation [Scholkopf et al., 1997].

The possibility of linearising the feature space and analysing the obtained support vectors make the SVM attractive from an interpretability point of view. Nevertheless, it is worth remembering that (as with any other classifier) if the number of features becomes too large, interpretability might be lost.

CNNs

CNNs are DL models originally built for image classification and thus, handle input positions explicitly. In addition, by transforming the image RGB property into a time series one, it is possible to handle the time dimension explicitly. It is also common to transform the time series raw data into a spectrogram through the Fast Fourier Transform (FFT) or wavelet decomposition [Usman et al., 2021b, Zhang et al., 2019, Truong et al., 2019, Truong et al., 2018]. Based on this, authors have the opportunity to handle EEG seizure prediction as a more powerful visual time series while accounting for electrodes positions and their relation [Xu et al., 2020, Sun et al., 2018, Khan et al., 2017].

The architectures of these networks can vary. Authors typically stack several convolutional layers, as these build feature maps through kernel filtering operations. These layers are then followed by pooling ones, responsible for learning features from the previously obtained maps. After these layers, it is possible to use classification layers. Dropout layers are also crucial as they prevent overfitting since these networks may overfit the training data due to their significant number of parameters. These layers set the output of random units to zero during training [Khan et al., 2017].

LSTMs

LSTMs are a type of Recurrent Neural Networks (RNNs) comprised of special units, named gates, that control (by learning the respective weights) input, output, and recurrent connections. As LSTMs have internal memory accumulated through sequential input and analysis, their analysis does not rely on a fixed window, which is an advantage when compared to CNNs [Sun et al., 2018]. Again, as DL models, these can be difficult to interpret and are prone to overfitting, thus requiring large amounts of data [Daoud and Bayoumi, 2019, Tsiouris et al., 2018, Schirrmeister et al., 2017].

3.1.6 Performance assessment

Lastly, a developed methodology must be evaluated according to a given set of metrics. Despite the seizure prediction characteristic [Winterhalder et al., 2003] advises using seizure sensitivity, False Prediction Rate per hour (FPR/h), and statistical validation, such as surrogate analysis or unspecific random predictors, not all authors follow this strategy. Although using the Area Under the Curve (AUC) [Stirling et al., 2021b, Nasseri et al., 2021, Truong et al., 2019], sample specificity, and sample sensitivity [Usman et al., 2021b, Assi et al., 2015] may provide valuable information concerning the classifier performance, these may not be adequate to understand how the correspondent system would fit into real-life where an intervention might be expected. Other authors [Viana et al., 2022, Pal Attia et al., 2022, Daoud and

Bayoumi, 2019, Cook et al., 2013, Karoly et al., 2017] shift from prediction to forecasting, where in that case, it is plausible the use of the Time in Warning (TiW) metric instead of presenting the FPR/h.

It is possible to observe that performance is severely influenced by the used database, where greatest heterogeneity in results can be observed for the EPILEPSIAE database [Direito et al., 2017, Rasekhi et al., 2015, Alvarado-Rojas et al., 2014,Teixeira et al., 2014b]. This variance may occur due to the considerable number of patients in this database (278). Authors typically make a patient selection based on the signal quality (low presence of noise and artefacts), epilepsy type, number of seizures, and lead seizure criteria. Patient number in these studies ranges from 1 [Rabbi et al., 2013] to 278 patients [Teixeira et al., 2014b]. CHB-MIT appears to be the database with the highest performance and homogenous results [Usman et al., 2021b, Xu et al., 2020, Zhang et al., 2019] and thus, perhaps making questionable its use to demonstrate predictive power of new methods. Nevertheless, contrarily to EPILEPSIAE and Neurovista databases, CHB-MIT is a free and open-access database.

Ultralong-term databases, such as the Neurovista [Kiral-Kornek et al., 2018, Karoly et al., 2017, Cook et al., 2013], Neuropace-derived [Nasseri et al., 2021], and others [Stirling et al., 2021b] appear to be the ones that might bring more realistic performances and thus, worth analysing towards the possibility of one day developing a commercial prediction device.

3.2 Concept drifts

The influence of Concept Drifts (CDs) in seizure prediction has been studied over the years. As mentioned in chapter 2, medication withdrawal in presurgical monitoring, circadian and sleep-wake cycles, implanting a neurostimulation device, or a sudden brain lesion may influence cerebral dynamics. Different strategies can be considered to handle each type of CDs. A gradual or reoccurring drift might be handled by training several classifiers, one for each concept. For incremental drifts, one might retrain a classifier over time. Lastly, there is also the possibility to use concept states (such as sleep-wake state, time of the day, and others) as additional features [Dümpelmann, 2019].

Authors have dedicated efforts to identify cyclic patterns related to seizure occurrence in different temporal scales: circadian, multidien, and circannual. Firstly, these cycles have been explored by assessing the seizure frequency obtained from patient seizure diaries. Depending on the patient, there is a higher seizure occurrence within certain times of the day, month, and year. Then, by analysing ultralong term recordings, patterns have been identified [Rao et al., 2021,Karoly et al., 2021]. Table 3.5 shows the results for the most important studies on seizure cycles.

Seizure diaries have been used to capture patterns of seizure cycles, where the

Table 3.5: Studies on seizure occurrence cycles.

Study	Patient data	Seizure cycle prevalence
[Leguia et al., 2021]	Analysing only EEG Seizures: 85 patients using the RNS System; Analysing both EEG Seizures and diaries: 186 patients using the RNS System; Analysing Seizure diaries only: 194 patients	Circadian: 89% of patients; Multidien: 60% of patients; Circannual: 12% of patients
[Baud et al., 2018]	EEG seizures: 37 patients using the RNS System	Circadian in: 86% of patients; Multidien in: 93% of patients; Seizures often occur during the rising phase of multidien interictal epileptiform activity rhythms
[Karoly et al., 2018]	EEG Seizures: 12 patients from NeuroVista study (during 2 years); Seizure diaries 1118 patients from SeizureTracker (during 9 years)	Circadian in at least: 80% of SeizureTracker patients and 92% of Neurovista patients; Circaseptan: between 7% and 21% of patients
[Ferastraoaru et al., 2018]	Seizure diaries: 10186 patients (up to during 8 years) from SeizureTracker	Circadian pattern: higher seizure frequency between 7am and 10am, and lower overnight. Multidien pattern: higher seizure frequency during the work days comparing to the weekend

simplest strategy consists in analysing patient diaries. Nowadays, there is the possibility to use online diaries, such as the SeizureTracker [Fisher et al., 2012] and mobile applications. However, one must consider that those seizure diaries made by patients and caregivers must be carefully interpreted, as these are subjected to natural limitations and biases. Consequently, chronic-EEG systems may significantly improve the number of identified seizures. In fact, the NeuroVista study [Cook et al., 2013] showed significant differences concerning electroencephalographic seizures and annotated seizures in patient diaries [Brinkmann et al., 2021, Rao et al., 2021].

By using long-term EEG signals, authors verified a circadian cycle influence ranging from 86% to 92% of patients [Karoly et al., 2018, Baud et al., 2018, Leguia et al., 2021]. The sleep-wake cycle has also proven to model seizure cycles, as there are peaks related to sleep-wake transitions. There may exist a mutual influence between circadian cycles and the sleep-wake states [Rao et al., 2021]. It is relevant to note that seizures do not occur in all circadian cycles. These cycles do not necessarily lead to seizures but rather increase their occurrence probability [Karoly et al., 2021].

Interictal epileptiform activity may also influence seizures. This patient-specific cycle has a multidien periodicity, where seizures occur during the rising phase. This activity can be, in part, be modulated by circadian and sleep-wake cycles [Karoly et al., 2021, Khan et al., 2018, Baud et al., 2018].

The information above can be used to improve prediction performance. [Schelter et al., 2006] analysed the obtained false predictions for both awake and sleep states in two prediction models (one trained with dynamic similarity index and another with mean phase coherence). They verified that 86% and 68% of false predictions occurred during sleep. [Karoly et al., 2017] compared the results of an electrocorticography-based logistic regression model, a circadian probability, and a combined electrocorticography and circadian model. The addition of circadian

information (the combined model) maximised performance across different metrics.

3.3 Explainability

A Google Scholar search was conducted using the following terms: "explainability" or "interpretability" along with "EEG seizure prediction", "EEG epilepsy", or "EEG". A reduced number of explainability studies concerning seizure prediction, seizure detection or classification, and other EEG-related tasks were also included. The majority of studies opted for DL models. Table 3.6 summarises the obtained results.

Attention mechanisms are one of the most used strategies [Phan et al., 2022, Priyasad et al., 2021,Baghdadi et al., 2021,Briden and Norouzi, 2021,Mansour et al., 2020], which aim at reproducing the concept of cognitive attention. Summarily, by focusing on few but relevant objects, cognitive attention allows humans to position themselves towards relevant stimuli and respond to them. In the DL case, authors include attention mechanisms as components of the developed network architecture, which focus more on the small but essential part of the data. While some authors

Table 3.6: An overview of explainability studies using the EEG signal, mainly for epilepsy-related tasks.

Study	Task	Classifier	Explainability Methods
[Moghaddam et al., 2022]	Seizure Prediction	SVM	Spatial coherence
[Lo Giudice et al., 2022]	Classification of seizures into Epileptic or Psychogenic Non-Epileptic	CNN	Permutation Entropy
[Phan et al., 2022]	Sleep stage identification	Seq2Seq Model (RNN)	Attention scores
[Dissanayake et al., 2021b]	Seizure Prediction	GDL	Graph visualisation
[Tang et al., 2021]	Seizure Detection and Classification	GNN	Occlusion maps
[Dissanayake et al., 2021a]	Seizure Prediction	CNN Siamese Networks	SHAP
[Priyasad et al., 2021]	Seizure Classification	Attentive fusion model with CNN	Attention mechanisms
[Baghdadi et al., 2021]	Seizure Detection and Classification	Attention-based Deep LSTM	Attention mechanisms
[Naze et al., 2021]	Classification between tonic-clonic and absence seizures	SVM (linear and RBF) RF and DT	Feature importance
[Gabeff et al., 2021]	Seizure Detection	CNN	Gradient ascendent SHAP
[Briden and Norouzi, 2021]	Classifying anxiety levels	Squeeze-and-Excitation Network	Attention Scores
[Lee et al., 2020]	Sleep stage identification	CNN	Extract representative patterns
[Zhang et al., 2020]	Seizure Detection	Adversarial learning framework	Attention-based CNN
[Mansour et al., 2020]	Seizure Detection	CNN, Bi-LSTM, Attention Network and FC layers	Attention mechanisms
[Thomas et al., 2020]	Seizure Detection	Bottleneck Network Architecture	Analysis of latent features
[Uyttenhove et al., 2020]	Detection of Epilepsy	t-VGG Network	Grad-CAM
[Hossain et al., 2019]	Seizure Detection	CNN	Network correlation maps (visualisation)
[Vilamala et al., 2017]	Automated Sleep Stage Scoring	VGGNet using transfer learning	Sensitivity maps
[Schirrmeister et al., 2017]	EEG Decoding and Visualisation	ConvNets	Network correlation maps (visualisation)
[Wang et al., 2017]	Detection of Epilepsy	SVM, RF, C4.5, SVM+RF, SVM+C4.5	Rule-based explanations

GDL stands for Geometric Deep Learning, GNN for Graph Neural Networks, RNN for Recurrent Neural Networks, RF for Random Forest, DT for Decision Tree, FC for Fully Connected, and VGG for Visual Geometry Group.

use these strategies to obtain the most relevant brain regions [Priyasad et al., 2021, Baghdadi et al., 2021, Zhang et al., 2020], others find the most crucial features [Mansour et al., 2020].

Other techniques, such as sensitivity maps [Vilamala et al., 2017] and occlusion maps [Tang et al., 2021] are also used to highlight the most discriminative parts of the input. Occlusion maps are developed by systematically occluding different portions of the input image and monitoring the classifier output. When objects are important, their occlusion in the input significantly lowers the probability of the correct class obtained in the softmax layer. Gradient-weighted Class Activation Mapping (Grad-CAM) [Uyttenhove et al., 2020] is also a possibility, which explores the last convolutional layer to weight the relevance of each neuron. The final output is a heatmap highlighting the input data that positively influenced the classification.

SHapley Additive exPlanations (SHAP) values are also common. They are designed to understand how the input is related to the output, namely the importance of given input parts/features [Gabeff et al., 2021]. It is also possible to use this strategy to determine channel importance [Dissanayake et al., 2021a]. Other ways to determine features' significance are permutation feature importance [Naze et al., 2021] and through permutation entropy [Lo Giudice et al., 2022].

Some authors develop graphic explanations which attempt to unravel hidden patterns in the brain, namely connectivity measures [Moghaddam et al., 2022, Dissanayake et al., 2021b]. Others correlated the obtained patterns with the spatial distribution of band power features as these may have an intuitive meaning for clinical purposes [Thomas et al., 2020, Hossain et al., 2019, Schirrmeister et al., 2017]. There is also the possibility of using clinically characteristic waveform shapes to pre-train the networks to reinforce the use of distinctive patterns [Lee et al., 2020].

Lastly, [Wang et al., 2017] employed rule-based classifiers. More specifically, an ensemble learning approach extracted a set of human-comprehensible rules from an SVM model, providing explanations for the model's decisions.

3.4 Final reflections

A typical seizure prediction study comprises the steps of signal acquisition, preprocessing, feature extraction, feature selection/reduction, classification, regularisation, and performance evaluation. These steps are performed through sliding window analysis, where the classification commonly concerns a binary classification model that distinguishes preictal from interictal EEG segments. The introduction of DL allows the possibility of automatically performing feature engineering and preprocessing from raw data while handling time dependencies intrinsically. These models have a higher potential for achieving high performance. However, they can be notoriously tricky to interpret, especially when compared with traditional ML models fed with a reduced number of features.

Despite recent advances in signal acquisition and ultra-long-term recordings, several authors still use datasets acquired in presurgical monitoring, which concerns a few weeks of data per patient. This may happen as a substantial quantity of ultra-long-term continuous data is not publicly available and presurgical evaluation needs to be performed for incoming Drug-Resistant Epilepsy (DRE) patients.

Feature engineering tasks present substantial heterogeneity among studies where several measures and complex classifiers may obtain higher performance at the expense of losing interpretability. Linear univariate features are one of the most used as they are fast to compute. It may be necessary to develop efforts to explain model decisions to clinicians in a more accessible way. Current efforts are mainly directed to DL approaches, which attempt to provide the importance of brain region (channels), patterns of brain connectivity, and correlations with band-waves activit, which may be intuitive to clinicians.

Despite the preictal period varies between patients and seizures from the same patients, authors have adopted fixed periods. It is defined as a point-of-no-return concerning a seizure, where an optimal duration has not been found. As the brain may have mechanisms to stop seizure generation processes, it may be more suited to define a seizure susceptibility state than a point-of-no-return one. Consequently, a shift from prediction to seizure forecasting has been observed over recent years. This vision is also more suited to handle recent findings regarding seizure occurrence cycles and the influence of circadian, multidien, and circannual CDs.

Chapter 4

Seizure Prediction Ecosystem

This chapter is a sociological study of the seizure prediction research field. Section 4.1 presents the study context. Section 4.2 details the used methods and materials. Section 4.3 provides the results in the form of a social network and research guidelines. Lastly, section 4.4 discusses the obtained findings and limitations, and provides final reflections. The content of this chapter is based on a journal article published in *Epilepsia Open* [Pinto et al., 2021b].

4.1 Study context

Although seizure prediction research started in the 1970s through Electroencephalogram (EEG) analysis [Gadhoumi et al., 2016b, Mormann et al., 2007, Iasemidis, 2003], few predictive devices [Cook et al., 2013] and closed-loop systems [Sun and Morrell, 2014] have been clinically approved for clinical trial. Additionally, these were based on the "detection of features alone" (line-length, bandpass, and energy-related) [Freestone et al., 2017], which may be less robust than current state-of-the-art approaches [Gadhoumi et al., 2016a]. An overview of current research uncovers the existence of major multidisciplinary barriers [Kuhlmann et al., 2018b,Kuhlmann et al., 2018b,Gadhoumi et al., 2016a]. For instance, to develop a trustful, robust, and commercial solution, one needs to handle expectations and beliefs from all actors of this ecosystem: technology and data scientists, clinicians, industry, legislation, ethics, and patients [Kuhlmann et al., 2018b, Goodman and Flaxman, 2017, Ramgopal et al., 2014, Schulze-Bonhage et al., 2010].

This chapter inspects the seizure prediction literature to understand the social difficulties, where this analysis was based on Grounded Theory (GT) [Chapman et al., 2015] and Actor-Network Theory [Cresswell et al., 2010]. GT is a standard methodology applied in qualitative research where researchers draw hypotheses from data: unlike most quantitative methods, data collection is not part of a process to test a pre-existing hypothesis. Thus, GT's method consists in the identification and iterative refinement of relevant subjects from data [Chapman et al., 2015, Boyatzis,

1998]. Actor-Network Theory's main characteristics focus on inanimate entities and subsequent effects on social processes. Technology emerges from social interests and configures social interactions instead of handling technology as an external force. Actor-Network Theory can be useful for studying information technology implementations in healthcare settings [Cresswell et al., 2010].

A social network [Scott, 1988] that describes the relations between all actors is presented here. Encapsulation allowed the network to deliver a more general overview while deepening technical aspects that can be accessed individually. Furthermore, exploring this ecosystem helped unravel paths that may lead to a higher chance of clinical acceptance. Trust plays a fundamental role in increasing the number of clinically approved studies and subsequent commercial devices. The absence of an explanation for black-box decision models, especially when they fail, makes researchers question and mistrust their use, thus raising scepticism. This is why some authors argue for using only interpretable models [Rudin, 2019].

However, for the specific case of seizure prediction, the obtained findings argue that efforts should focus on explainability (and not necessarily on intrinsically interpretable models) as it is sufficient to reinforce trust, patient safety, ethics, and compliance with applicable law and industry standards. Explainability may be the critical aspect that allows the entrance of promising Deep Learning (DL) approaches in clinical practice, as these hold great potential. Note that, as mentioned in the Background chapter (section 2.6), interpretability and explainability are different concepts [Gilpin et al., 2018]. While the former regards the extent to which a system output can be predicted by a given input, which is clear by using intrinsically interpretable models with a reduced set of features, explainability concerns how to explain the decisions made.

By providing a social understanding and guidelines for effective communication between actors, this work contributes toward new clinically trusted methodologies, particularly for the work of those who develop software seizure prediction approaches, so that they have a higher chance of clinical acceptance. Conversely, it may also help clinicians to understand this research area. Although the academic community may have implicitly used these guidelines for several years, their formalisation may be interesting and valuable. Many concepts developed here may also be applied to other healthcare areas where built devices implement algorithms developed from clinical data.

4.2 Materials and methods

The methodology comprises five stages (see Figure 4.1). Firstly, the selection of studies from the literature. This selection considered relevant studies that addressed seizure prediction models, patients' points of view, legislation, and algorithm explainability.

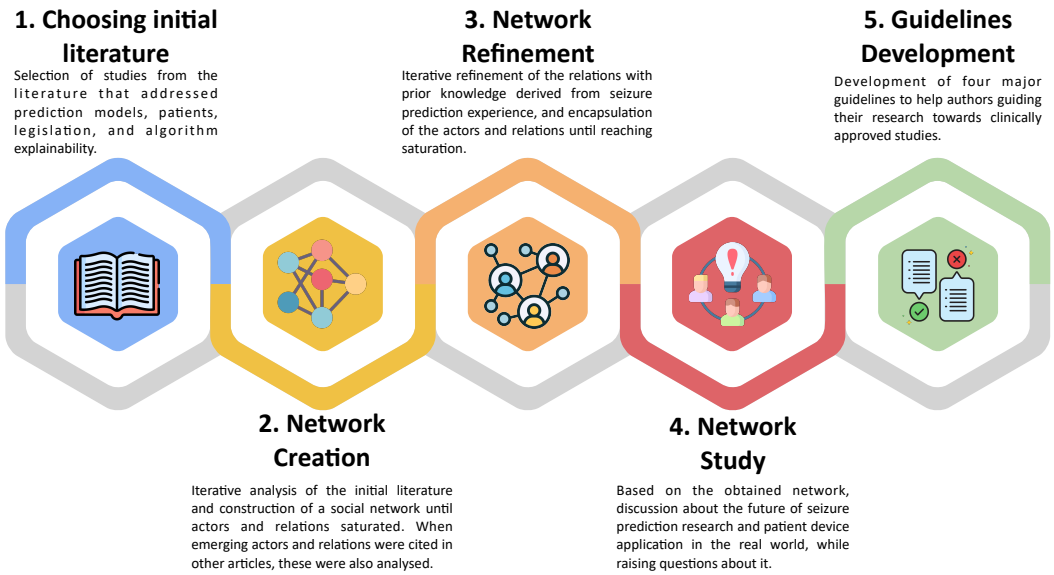


Figure 4.1: The five-stage methodology followed in this work. Icons obtained from [Freepik, 2021a, Freepik, 2021b, Icon, 2021, Becris, 2021a, Becris, 2021b].

Secondly, based on the latter, the development of a social network model occurred until reaching saturation or, more specifically, until there was not possible to find more actors or relations. Additionally, when actors and relations emerged from referenced studies in the selected papers, those were inspected to certify that saturation occurred. Thirdly, the social network was refined with prior knowledge derived from seizure prediction. Fourthly, the obtained network was studied to discuss the future of seizure prediction and possible devices for patients. In this discussion, several questions were listed. Other topics and studies also arose in this discussion with all authors and reviewers, which led to select more papers (see the Appendix C to understand how all papers were selected). Finally, developing four guidelines was crucial for the faster progress toward new clinically accepted studies.

4.2.1 Choosing initial literature

The starting materials were the published literature on seizure prediction, as this research field has almost 46 years of existence. Three surveys [Kuhlmann et al., 2018b, Freestone et al., 2017, Mormann et al., 2007] were chosen, which provided an overall vision of past, present, and future of seizure prediction. These present a critical view of the area. Additionally, a survey [Ramgopal et al., 2014] on seizure detection and prediction devices, and an article presenting Drug-Resistant Epilepsy (DRE) patients' view on seizure intervention devices [Schulze-Bonhage et al., 2010] were also selected. Finally, a book on interpretable Machine Learning (ML) [Molnar, 2019], available online, was chosen. This last selection is due to a prior awareness of the importance of interpretability/explainability. These materials were analysed in the order they are referenced in this paragraph. From all the stages, this is the one

(choosing the initial literature) that may lead to greater discussion among seizure prediction experts.

4.2.2 Network creation

A social network was created because it provides a power model for social structure. The concept of a network here is a set of points (actors, which can be individuals or collective) connected by lines (relations). The goal was to describe these relations and explain the patterns found. Constructing a network is not a theoretically-neutral task, as it depends on the intellectual judgement of the researcher [Scott, 1988]. The literature analysis was based on GT and Actor-Network Theory to help structure the network development. Some of the iterations of the developed network are found in Appendix D.

GT is an inductive process that has contributed to a broad acceptance of qualitative methods in several social sciences [Chapman et al., 2015]. Its fundamental premise is that researchers must develop a theory from empirical data. Its overall process consists of the codification of gathered data and identification of emerging themes, and consequent development throughout further data collection [Charmaz and Belgrave, 2007]. Coded data are commonly short statements or words that capture the meaning of phrases and are used to index data and group ideas.

Additionally, concepts of Actor-Network Theory were simultaneously used with GT. Thus, GT was not used to search traditional themes but rather to search for socio-technical actors and their relations. With these, a social network was built. The GT analysis was iterative and performed until reaching saturation. More particularly, it stopped when new actors or relations were not found [Chapman et al., 2015].

Although GT develops theories from rigorous data gathering, the research process requires a certain sensitivity [Davey and Adamopoulos, 2016, Chapman et al., 2015]. It is relevant to stress that the researcher's experience heavily influences the data codification and the emergence of themes and ideas. Therefore, the main criticism of this theory is the possible introduction of bias, given that truly inductive analysis may not be achievable. This work is limited by prior knowledge. Due to this, note the existing background experience in developing ML pipelines for healthcare, particularly in seizure prediction [Chapman et al., 2015].

Actor-Network Theory is a sociological approach to understand humans and their interaction with technology in specific settings. Its main characteristic is symmetry, which treats equally human and non-human objects [Troshani and Wickramasinghe, 2014]. It is a framework based on the following concepts [Cresswell et al., 2010, Wickramasinghe et al., 2007]: i) actors, the participants in the network which are human and non-human objects; ii) heterogeneity, each actor's importance is given by the web of relations; iii) quasi-objects, the successful outcomes which pass from actor

to actor within the network; iv) punctualisation, a similar concept to abstraction in object-oriented programming, referred here as encapsulation; v) obligatory passage point, situations that have to occur for all actors to satisfy the interests of the network; and vi) irreversibility, wherein healthcare is not likely to occur due to the importance of developing robust and effective studies to maintain patient safety.

At its heart, Actor-Network Theory tackles the notion of an organisational identity [Wickramasinghe et al., 2007]. It was used to guide this analysis to investigate, understand, and explain the processes that influence and lead to the development of clinically approved studies for seizure prediction [Iyamu and Mgudlwa, 2018]. Some criticisms [Cresswell et al., 2010] on Actor-Network Theory are that it may be too descriptive. Moreover, it fails to deliver any definitive explanation or approach that best handles the studied actors and relations. Another limitation is that it fails to handle human intentions, morals, backgrounds, and previous experiences of human actors. This was one of the reasons why highlighting the importance of explainability. A given explanation will depend on these, and, although intentions, morals, and backgrounds were not tackled directly, rigorous explainability evaluation on the application and human levels might account for them (as performed in chapter 6).

4.2.3 Network refinement

After the social network reached saturation, a complex structure with many actors and relations was obtained. The network could not be delivered in that form as it was not intuitive. The network was refined based on prior seizure prediction experience. This process was also motivated by the mentioned dependence on researcher sensitivity and punctualisation (encapsulation). It is believed that the existing inexperience in social sciences could have derived some of these problems. These could have been overcome differently by experienced researchers in social sciences, as they have a higher understanding of Actor-Network Theory stages [Wickramasinghe et al., 2007], such as inscription, translation, and framing.

As previously stated in this paper, certain relations were redefined, such as those concerning brain assumptions, confounding factors, performance, and trust. These refinements were performed until reaching saturation. Actors were also grouped in colours concerning themes found to be intuitive: signal acquisition and life-related (blue), studies (orange), people and exchanging beliefs (yellow), prospective applications (green), and brain dynamics that trigger seizures and how to capture its data (red).

4.2.4 Network study

Then, the network was discussed to make it robust and detect possible conflicts, irregularities, and missing actors/relations. Note that Actor-Network Theory investigates the description of the relations, how a network comes to being, and how it

temporarily holds. The addition or removal of an actor significantly affects the network. Thus, it may fail when dealing with changes by focusing on a stable situation.

As the seizure prediction experience from authors contributes to this work, the outcome might differ among researchers. Others may include different initial articles and perform differently on data codification, network refinement, and encapsulation. Additionally, it is relevant to remember that the network is permanently evolving as the social reality is constantly changing and is complex [Cresswell et al., 2010].

The assumptions made on brain dynamics were also discussed until reaching a consensus due to their particular importance. Finally, based on the social network, a discussion on probable paths for seizure prediction future was made, where several questions arose.

4.2.5 Guidelines development

At last, four consensual guidelines were developed that may lead to progress in this area. These were based on the obtained network, its development, and seizure prediction future discussion.

4.2.6 Interactive presentation

In the end, an interactive presentation was developed, which can be accessed in `https://onlinelibrary.wiley.com/doi/full/10.1002/epi4.12597` in Supporting Information. It allows the reader to explore the ecosystem and understand better the network's encapsulation. A simplified version of a seizure prediction product process is also presented there, from presurgical monitoring acquisition until prospective application development. Also, the reader is allowed to explore the whole ecosystem interactively.

4.3 Results

This section presents a summarised version of the seizure prediction ecosystem and the proposed guidelines. In Appendix E, the social network is provided in full detail. It is relevant to note that encapsulation aspects, other details, and a step-by-step product design explanation are presented more intuitively in the interactive presentation. The reader is allowed to explore the whole ecosystem interactively. This section also focuses on the findings related to clinical trials, explainability, and interpretability.

4.3.1 Seizure prediction ecosystem

Figure 4.2 depicts the obtained social network, which describes the relations between actors. Actors (x) and relations (x-y) are named with numbers and grouped in colours to provide a better understanding. This section will explain these relations

while deepening parts that require more detail. In the end, this section provides guidelines to help authors design their research.

The ecosystem begins with a DRE patient (1). Years after being diagnosed with DRE, a patient is referred to an epilepsy centre to undergo presurgical monitoring (5). The EEG signal (4) is acquired to inspect brain activity to localise the epileptic focus. If easily localised, removing the epileptic region is a possible solution [Engel, 2016, Mormann et al., 2007]. This data will be stored and constitute retrospec-

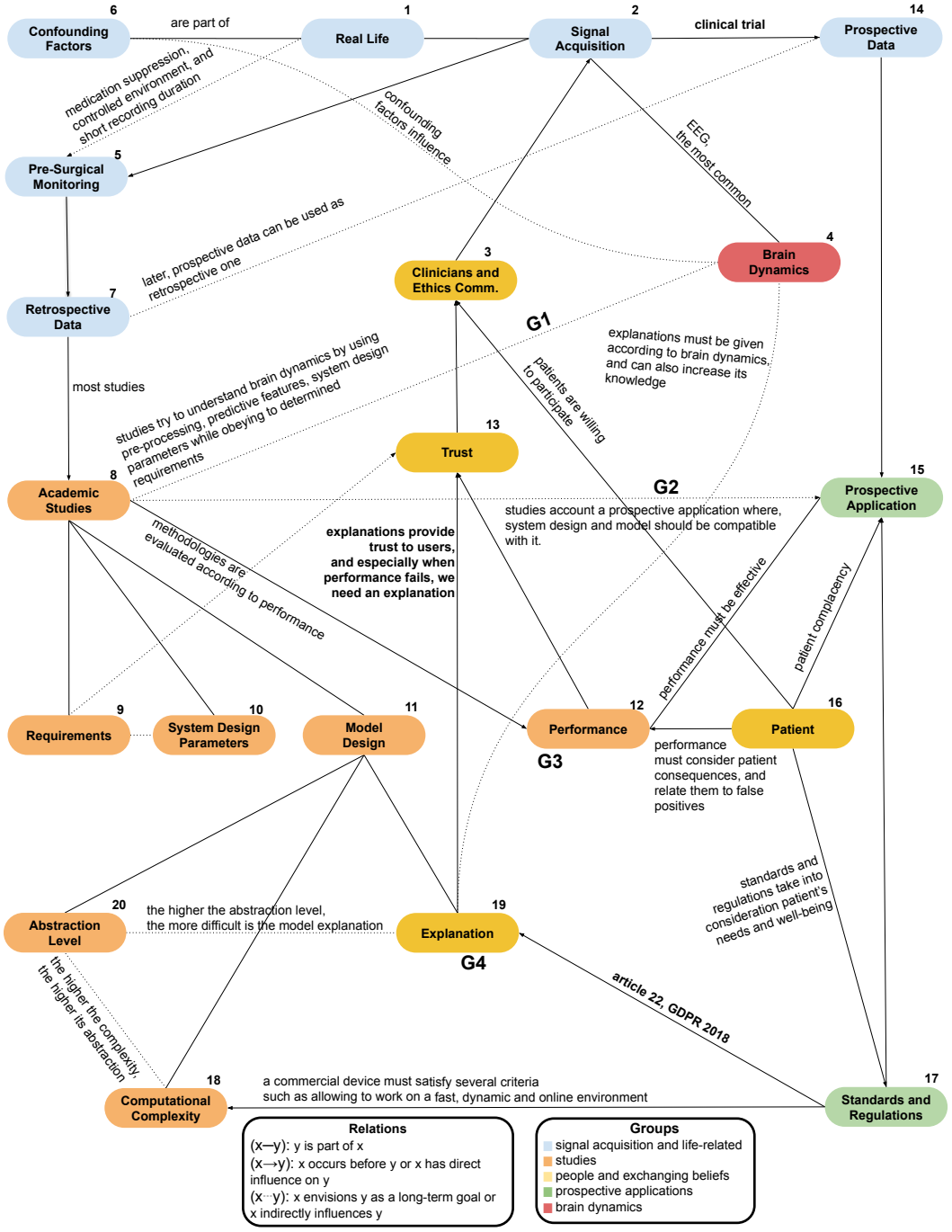


Figure 4.2: The obtained network.

tive data (7). The majority of databases available to perform academic studies (8) concerns presurgical monitoring conditions.

Studies try to capture and understand brain dynamics to predict seizures (8- -4). Inevitably, researchers make several assumptions (Appendix F for more information) when designing a new study. These may result from the used mathematical models, available data and other limitations, or even reflect the researcher's knowledge concerning brain dynamics (8- -4). These studies must also envision a real application scenario by simulating a prospective scenario (8- -15). Thus, studies must then comply with some requirements (9), have appropriate design parameters (10) concerning the real application, propose a discriminative model (11), and discuss its performance (12). Model design (19) is one of the most explored topics (which includes here preprocessing, feature extraction, and model training). A model can be characterised according to computational complexity (18) and abstraction level (20).

To start a clinical trial, trust (13) is necessary. Data scientists and clinicians need to find a given methodology trustworthy. There is a need to ensure patient safety, model robustness, and avoid bias. High performance is a necessary condition (12→13), but it is not enough. It is also needed to explain the developed model's decisions (19→13) to ensure safety and model effectiveness. Note that, for the particular case of seizure prediction, although it is required to know how to explain the model's decision, it may not be necessary to use intrinsically interpretable models, as seen in the following sections with the Neurovista Advisory System [Cook et al., 2013].

For clinical trials, this work argues the possibility of using complex prediction models, including black-box systems to some extent, as long as authors provide efforts to avoid data bias, ensure patient safety, and explain their models' decisions. Furthermore, explanations not only increase trust and mitigate scepticism on artificial intelligence algorithms, but they can also deliver new knowledge on brain dynamics (19- -4).

Concerning legislation (17), the 2018 General Data Protection Regulation (GDPR) also promotes the delivery of model explanations (not necessarily intrinsically interpretable models). Current legislation should be seen as a reinforcement of safe methodologies that considers patients' needs and well-being (13→17). When data scientists and clinicians trust the proposed methodology, the ethics committee can accept a clinical trial (13→3). Patients are invited to participate in clinical trials (16→3) in this case.

After the ethics committee's approval and patients' agreement to volunteer, a clinical trial starts. The prospective data (14) later becomes retrospective (7) and is used in an indefinite number of studies. During the acquisition of the prospective data and by timely anticipating seizures, it is possible to apply an intervention in real-time. To do this, researchers must guarantee that the false-positive interven-

tions are not harmful to the patient (16→15). The intervention must also comply with industry standards and safety measures (17). It must have fast processing, no hardware problems, and easy placement and removal.

4.3.2 Studies guidelines

By describing and discussing all relations, four guidelines were inferred that might help authors guide their research on seizure prediction. Figure 4.3 depicts a production process of a hypothetical device. Firstly, authors perform studies with retrospective data, evaluating performance and the quality of given explanations. Clinicians and data scientists trust models' decisions when these are human-comprehensible, also increasing the confidence of the volunteering patients. In this case, an ethics committee may have strong reasons to approve a prospective study with an intervention system. Finally, the built device reaches its goal: to improve the life of DRE patients.

The first guideline ($G1$) concerns undertaken assumptions on brain dynamics, which differ between studies due to available data and used methodology. Authors should state their assumptions regarding brain dynamics before presenting the mathematical tools used in data analysis. Experienced researchers may understand what is at stake. However, others may benefit from the assumption statement by gaining faster insight, enabling easier comparison among studies, and understanding limitations. For instance, authors claim that tackling confounding factors increases performance, but believing in a direct causal relation may be naive. Reducing confounding factors does not increase performance *per se* but rather improves the experimental design and study requirements by improving assumed brain dynamics

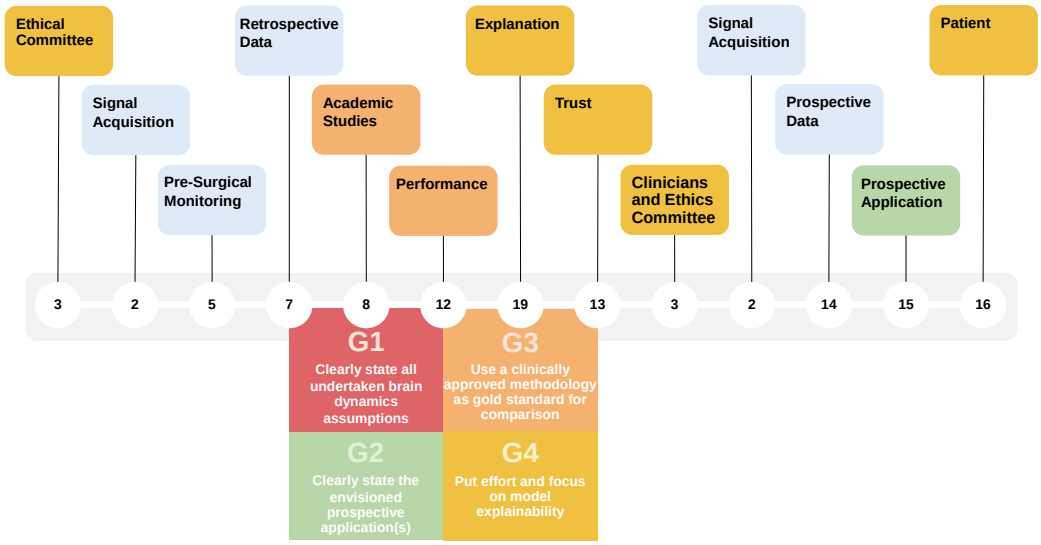


Figure 4.3: A product process for a seizure prediction prospective application, while showing the obtained guidelines concerning designing academic studies.

(8- -4), namely in model design and problem definition. Similarly to confounding factors, aspects such as problem definition and system design parameters encounter the same problem.

The second guideline ($G2$) concerns stating the prospective applications envisioned with the designed experiment (8- -15). It helps readers and authors understand what is at stake concerning system parameters, data type, and envisioned intervention. For instance, most seizure prediction studies report optimal Seizure Occurrence Period (SOP) periods for 30-60 minutes. Nevertheless, the RNS®system is programmed to make electrical discharges up to 5000 ms [Sun and Morrell, 2014]. For closed-loop systems, these SOP intervals are too long to deliver an effective intervention. Additionally, many authors use short Seizure Prediction Horizon (SPH) intervals in scalp EEG studies [Assi et al., 2017]. In these cases, an SPH of 10 seconds or even 1 minute is not enough for taking action after an alarm, such as reaching a secure place or taking rescue medication. For example, diazepam rectal gel (the only Food and Drug Administration (FDA) drug approved for seizure cluster and which might be tested as prevention) takes 5-10 minutes to work [Dreifuss et al., 1998]. In comparison, oral diazepam or lorazepam takes 15 minutes [Foundation, 2020]. This guideline would stimulate discussion regarding study limitations, as well.

The third guideline ($G3$) is related to using methodologies that have been clinically approved as a gold standard for comparison. Reporting only sensitivity, specificity, and prediction above chance level might be limited, as these metrics strongly depend on data and may not explicitly show progress. Authors should compare their approaches with the ones already clinically approved. This comparison should not only be based on performance but also explainability. The latter leads to the essential guideline ($G4$): researchers should focus on explainability (19) to promote trust among experts. It would be interesting to, at least, present a concrete example of model decisions throughout time. This way, it would demonstrate how a model could explain its predictions to an expert as a data scientist/clinician (application level) and a patient (human-level).

4.3.3 The importance of how explaining decisions

After a proper studies comparison, one can ask what a good performance is or even inquire about the minimum performance that justifies the design of a clinical trial. A proper methodology is one which researchers trust. In literature, trust seems to be represented by literature convergence and reproducibility, where studies report high performance (12→13) and comply with consensual study requirements (9- -13). By analysing data from longer recordings and/or a higher number of patients, trust increases as the testing data are more likely to represent real-life conditions [Kuhlmann et al., 2018b].

High-level abstraction models may potentially handle complex dynamics but re-

quire strong efforts toward providing explanations (19- -20). Current clinical knowledge of physiology should be the source of explanations and the basis for new findings (19- -4). As an explanation is an exchange of beliefs [Lombrozo, 2006], its acceptance may differ among patients, clinicians, and data scientists.

Although a given methodology eventually makes incorrect decisions, researchers can still trust it if one can explain its decisions (19→13). A great scepticism concerning ML and high-level abstraction models may be due to the difficulty in delivering explanations about models' decisions [Molnar, 2019]. Although authors and/or clinicians are more willing to trust black-box models when they make correct decisions, wrong ones lead to mistrust because there is no human-comprehensible explanation [Freestone et al., 2017].

The phase IV Neuropace RNS®system [Sun and Morrell, 2014] (NCT00572195) can use up to two independent detections, which are highly configurable and adjusted by the physician, which ensures patient safety. Each detection performs a threshold decision based on a given extracted feature (line-length, bandpass, and area) by comparing the current analysis window with another considered to have interictal activity. This is the most explainable and straightforward strategy that is possible to obtain. One can fully understand the underlying mechanisms behind each decision. The phase I NeuroVista Seizure Advisory System [Cook et al., 2013] (NCT01043406) is more complex, using a preprocessing step, extracting similar and intuitive features (line-length, Teager-Kaiser energy, and average energy), and training a ML model that produced a measure of seizure-risk which concerns a seizure-susceptibility state. This model uses as input the best 16 features (from a set of 16 channels X 6 filter/normalisation options X 3 analysis methods), and it involved ten layers (creating different decision surfaces), being inspired in k-Nearest Neighbours (kNN) and decision tree classifiers, where each layer considers a different seizure-risk related to its proximity to a seizure event. This algorithm is more complex and not fully transparent. In other words, researchers do not understand its underlying mechanisms, despite using kNN and decision tree classifiers (which may be intrinsically interpretable when using a reduced set of features). Calculating seizure risk in a 16-dimension feature space further divided into 2^{10} partitions (decision surfaces) is not human-comprehensible. Nevertheless, the extracted features are clinically intuitive, and the model decision can produce a human-intuitive output explanation of the obtained seizure risk. It simultaneously compares the current analysis window with several data distributions whose time proximity to a seizure (and therefore, seizure risk) is considered. Performing multiple data-distribution classifications may be more robust to data bias and noise. The authors also ensured patient safety: firstly, they accessed model performance on pre-acquired patient-specific data and secondly, only patients with satisfactory performance received the advisory system.

These two clinical trials demonstrate that, despite all the scientific community

efforts to develop complex models and consequent increase in performance, it may be necessary for a fully explainable model to provide trust. Additionally, the Seizure Advisory System clinical trial demonstrates the possibility of using models that are not necessarily intrinsically interpretable, as long as they produce human-comprehensible explanations while ensuring patient safety, handling data bias, and achieving model robustness.

4.4 Discussion

Despite useful for clinicians and patients to understand this ecosystem, this study is directed to researchers who develop prediction approaches to have a higher chance of clinical acceptance. Providing a comprehensible overview of all the ecosystems was difficult due to the author's data science/clinical background. Hence the natural bias/emphasis on academic studies. Although this work's limitations toward qualitative research tools have already been mentioned, it is relevant to stress its importance in the discussion as it constitutes a study limitation. For future work, interviews may be performed to provide possible paths and sub guidelines from the obtained ones. Appendix G presents a series of questions that arose from describing this ecosystem and may be desirable to tackle.

Regarding guidelines, $G1$ improves methodology comparison while delivering a deeper understanding of study limitations to clinicians (regarding assumptions on the underlying physiological mechanisms). For instance, it is interesting to note that, despite most authors with retrospective data using the preictal concept as a point of no return, the two clinically approved studies use seizure susceptibility instead, which shows potential for seizure forecasting. Forecasting is different from prediction, as it shifts away from whether a seizure will occur or not and focuses instead on identifying periods of a high probability of seizure occurrence [Dumanis et al., 2017]. Despite this study's particular emphasis on seizure prediction, these guidelines and conclusions can be adapted and, thus, hold for seizure forecasting (see Appendix H).

$G2$ increases the author's comprehension of the limitations of signal acquisition methods and patient consequences associated with the obtained specificity. Furthermore, increases in model performance at the cost of developing systems with unreal parameters may be questionable [Assi et al., 2017,Gadhoumi et al., 2016a]. Although large seizure occurrence windows may translate into higher performance, the interval to accept true alarms is larger. For the case of a warning system, researchers need to consider the levels of stress and anxiety induced on patients or the consequences of frequent intake of rescue medication [Gaínza-Lein et al., 2017, Scheepers et al., 2000, Tasker, 1998]. There is also the need to understand how/if closed-loops intervention systems can be used with significantly long occurrence periods [Sun and Morrell, 2014]. Authors develop methodologies that may lack practical applica-

tion by only considering an increase in performance as one of the primary research goals. Although some studies may have a primary goal of increasing knowledge on brain dynamics, researchers should clearly state limitations towards real application. Based on this, authors should be advised to study the patient consequences that arise from a given seizure intervention system by defining a maximum number of false alarms. For a warning device, the literature has pointed to a maximum of 0.15 in False Prediction Rate per hour (FPR/h) [Winterhalder et al., 2003] and a minimum of 90% sensitivity [Schulze-Bonhage et al., 2010]. For more details and to better understand what an acceptable performance for a clinical setting could be, see the Appendix I.

Legislation and industry standards can be understood as keepers of best practices on patient safety and trust among all actors. A holistic understanding of trust, explainability, and performance when developing a seizure prediction methodology may be crucial for this ecosystem. In 2007, [Mormann et al., 2007] declared that algorithms were still too limited in performance to justify enrolling in clinical trials using responsive stimulation. Although this paper is one of the most influential in seizure prediction, the first clinical trial (a warning system) [Cook et al., 2013] started only three years later, in March 2010 and was published in 2013. With this, it is possible to claim the following: despite some authors advocating performance limitations to justify clinical trials, these were performed in the past and continue to be. Thus, a limited performance to justify a clinical trial may be misleading. Clinical trials continue to be performed (as in the case of SeizeIT2, which ended in 2021) because researchers and ethical committees find them necessary, existing a favourable benefit/risk ratio. There is an ongoing necessity to perform clinical trials, especially to avoid publication bias. In the literature, it is easy to find prediction performances that are overestimated as authors, in some cases, only report the best results. When a methodology appears promising, it must be tested in different datasets and contexts.

Moreover, the first clinical trial using responsive stimulation (phase III RNS®System Pivotal Study, NCT00264810) started in 2005, leading to the phase IV clinical trial (RNS®System LTT study, NCT00572195) that started in 2006. All current-generation of clinically approved studies and intervention devices use the detection of features alone [Freestone et al., 2017], which demonstrates the importance of explainability. Other examples are present in the literature that arose during discussion as in 2014, [Teixeira et al., 2014a] tested the Brainatic, which is a real-time scalp EEG-based seizure prediction system, approved by the Clinical Ethical Committee at the Centro Hospitalar e Universitário de Coimbra. They computed 22 univariate features per electrode and used non-interpretable models, such as support vector machines, multilayer perceptron, and radial basis functions neural networks. Based on this, increased performance cannot be the single criteria for a favourable ethics committee decision. These examples show the existence of room for improvement,

possibly by exploring more complex but still explainable systems. For instance, the RNS®system might benefit from a more robust approach to capture dynamics before a point of no return [Mormann et al., 2007]. Towards this, more studies, such as the one by [Sisterson et al., 2020], need to be performed to assess the algorithm effectiveness of responsive neurostimulation. Conclusively, as these methods have been clinically accepted and since a gold-standard comparison method is missing, they should be used for performance comparison and decision explanation.

Computational power has increased in recent years, allowing DL approaches in several areas. Seizure prediction is no exception [Freestone et al., 2017, Nurse et al., 2016]. As these approaches, along with rigorous preprocessing [Freestone et al., 2017] have a higher potential to handle brain dynamics, and as intrinsically interpretable models may not be required to undergo a clinical trial, there is an urgent demand for developing explainability methods that work on top of black-box models.

There might be a tendency to argue that requiring an explanation will limit the model's performance (hypothetically 12→19). However, explanations may enhance the model's functioning by tackling the incompleteness of problem formalisation. In medical contexts, for example, a correct decision only solves the problem partially [Molnar, 2019, Doshi-Velez and Kim, 2017], which may also be context-dependent, as ethical issues may arise (e.g., choosing between saving a life and prolonging the suffering of a patient). Researchers want to simultaneously deepen brain dynamics understanding, detect data bias, and improve model robustness. Therefore, it is important to understand possible trade-offs between potentially related aspects that might not be easily recognised. When considered in an explanation, all of these improve understanding, which represents a way to promote patient safety and increases the chance of social acceptance concerning ML use [Doshi-Velez and Kim, 2017, Schirrmeister et al., 2017].

4.4.1 Translation to other healthcare problems

Despite being oriented to seizure prediction, obtained guidelines and relations may be easily translated to different healthcare problems. Other conditions may benefit from a real-life intervention, such as the case of deep brain stimulation in Parkinson's disease [Okun and Foote, 2010]. Computer-aided diagnosis/prognosis software tools face similar problems on ethics, explainability, and trust, given the high risk associated with model decisions in healthcare.

These guidelines and methodology can be applied to other healthcare settings using computer-assisted diagnosis/prognosis. However, guideline $G4$ may differ among situations. When predicting hospital mortality after acute coronary events, for example, there are established score models and, therefore, using intrinsically interpretable models might be required to integrate existing clinical knowledge [Granger et al., 2003] better. In the case of seizure prediction, obtaining interpretability can

become even more complex because (a) there is no clinical annotation on the preictal period [Mormann et al., 2007] and (b) the EEG is still far from being fully understood [Kuhlmann et al., 2018b, Freestone et al., 2017]. Therefore, it might be hard to replicate a methodology as there is no standardised protocol to identify the preictal period manually. When discussing case studies with clinicians on the EEG signal, it has been observed that they often tend to point to/annotate spikes-and-wave discharges, activity increase, and rapid changes in the signal morphology and associate these with seizure events or seizure susceptibility.

A possible way "to engage in the clinical discussion" would be by using complex models, such as Convolutional Neural Networkss (CNNs) to capture complex dynamics and then, by delivering (pointing) to the EEG-detected events that were considered for a given decision. This type of explanation could be performed by, for example, Local Interpretable Model Agnostic Explanations [Ribeiro et al., 2016], and should be evaluated beforehand at the application level of explainability by discussing these detected events with clinicians. This way, researchers might try to emulate the process of analysis of the EEG of an epileptic patient typically conducted by a clinician. Using such models may also unravel new patterns (EEG morphologies) that have not yet been associated with epileptic manifestations.

Indeed, the body can be envisioned as a black-box system. In the case of antidepressants, for example, there is still no explanation for the delayed effect of antidepressant drugs and what neurochemical changes reverse the many different symptoms of depression and anxiety [Harmer et al., 2017]. In short, researchers know the inputs (medication) and the outputs (the change in the patients). They do not fully understand the underlying mechanisms. These drugs are widely used because they are effective and have a favourable risk-benefit balance.

4.4.2 Study limitations

The most significant limitation was the patient role, as it did not properly include their agency. The academic community is still far from understanding what it is like to be a patient: the patients' expectations are largely different from those of clinicians and data scientists. There must be a stronger awareness of the active role that a patient can have in the future.

The case of Dana Lewis and Hugo Campos are examples where the patients might be able to track their data, analyse it, and, therefore, better control their closed-loop systems [Lewis, 2019, Chu et al., 2016]. Dana Lewis created the "Do-It-Yourself Pancreas System" (#DIYPS), founded the open-source artificial pancreas system movement (#OpenAPS), and advocates patient-centred, -driven, and -designed research. She created #DIYPS to make her continuous glucose monitor alarms louder and developed predictive algorithms to timely forecast necessary actions in the future (https://diyps.org/about/dana-lewis/). Hugo Campos was diagnosed with hyper-

trophic cardiomyopathy: a disease in which the heart muscle becomes abnormally thick, which can be fatal. He received an implantable defibrillator, a device that electrostimulates the heart in case of dangerous arrhythmias. After losing his health insurance, he bought a pacemaker programmer on eBay and learned how to use it with a two-week course. Hugo Campos is now a data liberation advocate and leader in the e-patient movement (https://medicinex.stanford.edu/citizen-campos/).

Article 22 of GDPR provides patients the right to have an explanation for any algorithm decision and to question those decisions. The complexity of these issues is worth noting, as this study presents an oversimplification. Patient accountability and its relation to clinical accountability will undoubtedly be discussed in the future.

4.4.3 Final reflections

The application of ML and the consequent requirements on interpretability/explainability depend on the context and available medical knowledge. For seizure prediction, this work argues the clinical use of DL if researchers put efforts into ensuring patient safety in each study and clinical trial stage. When ensuring a good risk-benefit balance for the patient and human-comprehensible explanations are provided, with patients willing to volunteer, it may even be unethical to limit new methodologies.

Future work should tackle the most relevant questions that arose during the previous stage by undergoing interviews with clinicians, data scientists, lawyers, and patients.

Chapter 5

Interpretable evolutionary algorithms

This chapter concerns the use of Evolutionary Algorithmss (EAs) to develop Machine Learning (ML) models that are easy to explain. This chapter also shows how to obtain knowledge on the brain that might be useful to clinicians. Although the previous chapter 4 concludes that intrinsically interpretable models may not be necessary to achieve trust and due to existing scepticism on black-box models, it is desired to understand their potential. Section 5.1 presents the study context. Section 5.2 details the used methods and materials. Section 5.3 provides the prediction results and how to extract brain knowledge. Lastly, section 5.4 discusses the obtained findings and limitations, and provides some final reflections.

The content of this chapter is based on two journal articles published in *Scientific Reports* [Pinto et al., 2022, Pinto et al., 2021a]. The presented methodology, discussion, and results are primarily based on the [Pinto et al., 2022] article. The 2022 paper can be interpreted as an extension of the [Pinto et al., 2021a] article. The code for both articles is available at the Github page https://github.com/MauroSilvaPinto.

5.1 Study context

The most common seizure prediction ML pipeline has two main limitations. Firstly, feature selection is commonly based on the discriminating power of each feature individually, or by using wrappers and embedded methods that address synergies but require a large computational power [Assi et al., 2017, Mormann et al., 2007]. Secondly, this framework is modular and composed of independent stages. Feature selection is usually not based on the final seizure prediction performance but rather on distinguishing preictal from interictal independent windows of fixed size. Therefore, the interaction between stages is not handled. Additionally, a fixed sub-set of electrodes and features are often considered at a given time instant, not allowing the

evaluation of lagged values of corresponding features, i.e., not considering temporal dynamics. Also, the choice of a fixed preictal period follows a grid search approach of different periods, e.g., 2, 20, 30, 60, or even 240 minutes [Moghim and Corne, 2014, Park et al., 2011].

More recently, Deep Learning models, such as Recurrent Neural Networks (RNNs), Long Short-Term Memorys (LSTMs) and Bi-LSTM, were introduced in seizure prediction [Khan et al., 2017, Mirowski et al., 2008]. Due to their underlying mechanisms, they are more suitable for time-series analysis than traditional classifiers. Despite the theoretical potential of these models to handle brain dynamics and the existence of notable efforts to retrieve interpretable insights (where the Electroencephalogram (EEG) signal is no exception [Schirrmeister et al., 2017, Schirrmeister et al., 2017]), some clinicians may not be willing to make high-stake decisions based on them [Rudin, 2019]. Low-complexity algorithms with interpretable insights (as the ones using intrinsically interpretable models), able to provide a deeper understanding of the ictogenesis process, may be favoured over others [Gagliano et al., 2019, Kuhlmann et al., 2018b, Freestone et al., 2017].

To tackle interpretability, synergy concerning features, and interaction between all pipeline stages, a solution may lie in constructing a search algorithm that selects a reduced set of computationally efficient and widely used features. This search algorithm should select features by looking at the pipeline as a whole and not as a sequence of independent stages. This work proposes an EA to handle this problem, as these algorithms have become effective for several tasks such as direct search, optimisation, and machine learning problems [Eiben and Smith, 2003]. They can be seen as population-based search algorithms that mimic natural evolution by evaluating individuals' quality through evolution operators (crossover and mutation) and a fitness function. A population is a group of individuals, where each one is represented by a point in the search space. The fittest individuals, evaluated by their fitness function values, tend to survive and propagate the genetic material by reproducing or mutating [Bartz-Beielstein et al., 2014, Eiben and Smith, 2003, Mitchell and Taylor, 1999].

This study concerns a patient-specific search algorithm aiming at seizure prediction while automatically trying to discover the preictal period based on evolutionary computation. Each individual in the EA population is a set of five features. Simply put, the set of features (individuals) that best perform in seizure prediction using a logistic regression classifier (fitness function) survive and proliferate, while the remaining die and do not contribute to propagating their genes, similarly to natural selection. From a technical ML point of view, this method uses the predictive power of a set of features and their synergy. From an interpretability point of view, it tries to provide a deeper understanding of the seizure generation processes by giving results that can be interpretable and by assessing gene interaction using the *apriori* algorithm, a classical association rule method [Borgelt and Kruse, 2002]. Patient

comfort was also studied by assessing the electrodes that provide discriminative information. Higher patient comfort was assumed when promoting solutions that do not require many electrodes (minimise the number of used electrodes) and focusing on a particular brain region (minimise the number of analysed lobes). Towards that end, the developed EA is multi-objective as it searches for the best trade-off between seizure prediction performance and patient comfort.

This methodology was tested, in a quasi-prospectively approach, using data from 93 patients from the EPILEPSIAE database [Klatt et al., 2012, Ihle et al., 2012] with several types of focal (temporal, frontal, occipital, and central) and generalised epilepsy. The envisioned real-life application concerns a patient-specific EEG scalp system with up to five electrodes to ensure patient comfort. Its preictal period ranges from 30 to 75 minutes, with an intervention time of 10 minutes for each prediction. It gives the patient sufficient time to avoid accidents and/or rescue medication intake. This methodology was compared with a control method, which is an adaptation from previous seizure prediction studies [Direito et al., 2017, Cook et al., 2013], to better compare the obtained performance.

5.2 Materials and methods

This work concerns a patient-specific approach, where the following procedure was applied to each patient: data preprocessing, feature extraction, training and testing, as depicted in Figure 5.1.

The raw EEG data of each patient was filtered and segmented into non-overlapping 5-second windows from which features were extracted. Next, the first three chronological seizures were used as input to the Multiobjective Evolutionary Algorithm (MOEA), followed by the selection of the individuals (set of five features and correspondent preictal period) from the Pareto-front (made of three objectives) with the best trade-off between objectives (sample sensitivity, sample specificity, and patient comfort). Then, these five MOEA output features and correspondent preictal periods were tested with the remaining seizures and compared with the seizure prediction method used as control. Since some parts of the MOEA are stochastic (initialisation and evolution operators), a different set of Pareto-optimal solutions can be obtained for each execution. Consequently, 30 MOEA executions were performed for each patient. Then, a phenotype study was performed for all patients to understand the predictors' decision mechanisms and infer about possible patients' seizure generation processes.

Concerning the MOEA output features, note that these were based on the concept of feature construction [Sondhi, 2009, Motoda and Liu, 2002, Liu and Motoda, 1998]. In this work, first-level features concern the ones directly extracted from the EEG. Then, the second-level ones, which constitute the MOEA phenotype, were computed by windowing and applying a mathematical operator to these features.

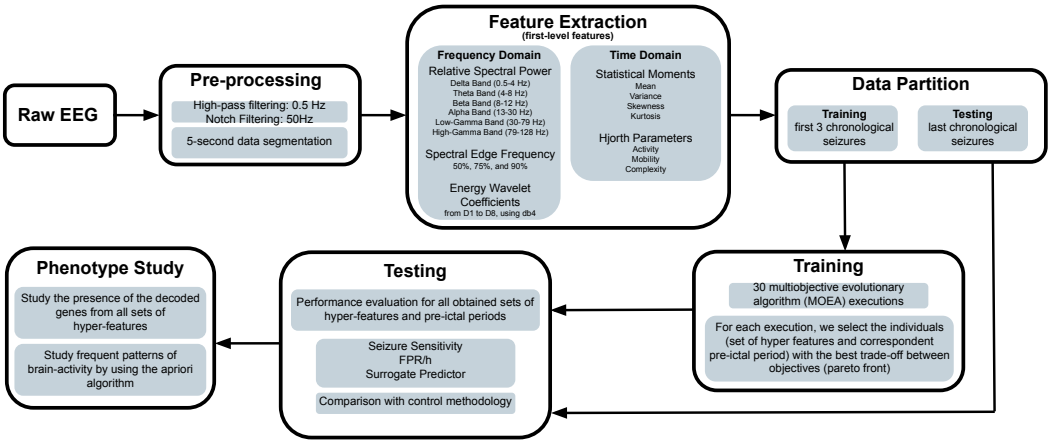


Figure 5.1: Flowchart of the proposed seizure prediction MOEA for each patient, comprising data processing, feature extraction, training, testing, performance evaluation and phenotype study.

Henceforth, first-level features are referred as *features* and second-level features as *hyper-features*.

5.2.1 Dataset

The patient selection criteria were: i) patients having a minimum of four lead seizures separated by periods of at least 4.5 hours; ii) patients with EEG scalp recordings, placed according to the 10-20 system and using a sampling frequency of at least 256 Hz; iii) patients with no more than one hour of EEG data missing for each seizure. All the analysed data were collected while patients were at the clinic under presurgical monitoring.

The use of this data for research purposes has been approved by the Ethical Committee of the three hospitals involved in the development of the database (Ethik-Kommission der Albert-Ludwigs-Universität, Freiburg; Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé, Pitié- Salpêtrière University Hospital; and Comité de Ética do Centro Hospitalar e Universitário de Coimbra). All methods were performed following the relevant guidelines and regulations. Informed written patient consent from all subjects and/or their legal guardian(s) was obtained to participate. Informed consent was obtained from all subjects and/or their legal guardian(s) for online open-access publication. Appendix J provides more information regarding the selected patients.

5.2.2 Preprocessing and feature extraction

All patient data were downsampled to 256 Hz, segmented into 5-second non-overlapping windows, and filtered with: i) a 50 Hz fourth-order notch filter (to remove the power-line interference) and ii) a fourth-order Butterworth high-pass filter with a cut-off frequency at 0.5 Hz (to remove the DC component and minimise motion artefacts).

Twenty-four linear univariate features were extracted for each time window in each electrode, 2. These are widely used in the literature [Bulusu et al., 2021, Assi et al., 2017,Direito et al., 2017,Gadhoumi et al., 2016b,Teixeira et al., 2014b,Rasekhi et al., 2013] and fast to compute. To understand each feature *a priori* expected added value, see Appendix B. It is also important to note that several features would be of interest [Stacey et al., 2020,Jacobs et al., 2018,Assi et al., 2017,Gadhoumi et al., 2016b,Bai et al., 2015,Kramer and Cash, 2012,Kramer et al., 2010]. These were not used as they are not univariate (it would be required to extract a considerably large number of features for each electrode) and are computationally heavier.

In the time domain, the first four statistical moments (mean, variance, skewness and kurtosis) and the three Hjorth parameters (activity, mobility and complexity) were extracted. As for the frequency domain, the following features were extracted: the relative spectral power of the delta (0.5-4Hz), theta (4-8Hz), alpha (8-12Hz), beta (13-30Hz), low-gamma (30-79Hz), and high-gamma (79-128Hz) bands, the spectral edge frequency at three different cut-off percentages (50%, 75%, and 90%), and the energy of each wavelet coefficient (D1 to A8, using the Daubechies 4 mother wavelet (db4)).

As the frequency limit of gamma activity is not consensual among the scientific community, and its division into high-gamma and low-gamma is not uncommon [Jia and Kohn, 2011], it was decided to divide it. Additionally, the gamma band powers may likely contain muscle artefacts as these recordings are extracranial. Therefore, gamma-band powers may not be fully considered as EEG features since these may represent physiological markers that predict seizures. Some authors [Bandarabadi et al., 2015b] report the difficulty of removing artefacts without eliminating good information; therefore, they may use raw signals.

5.2.3 Multi-objective evolutionary algorithm

Figure 5.2 depicts the employed MOEA, based on the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [Deb et al., 2002]. Firstly, the MOEA was randomly initialised with a fixed-size population of 100 individuals. Each individual (set of five hyper-features and a preictal period) is encoded by a chain of characters (genotype), which is further translated, from the problem context to the problem-solving space, leading to a possible solution (phenotype).

Then, each individual is evaluated according to the fitness functions, which are two measures related to seizure prediction performance (sample sensitivity and sample specificity) and another for patient comfort (based on the used electrodes and lobes). Based on the individuals' fitness values and their spatial spread, two steps are performed: i) ranking the individuals using non-dominated sorting and the crowding distance [Deb et al., 2002], and ii) choosing half of the individuals (parents) to reproduce (variation operators) by using binary tournament selection (parent selec-

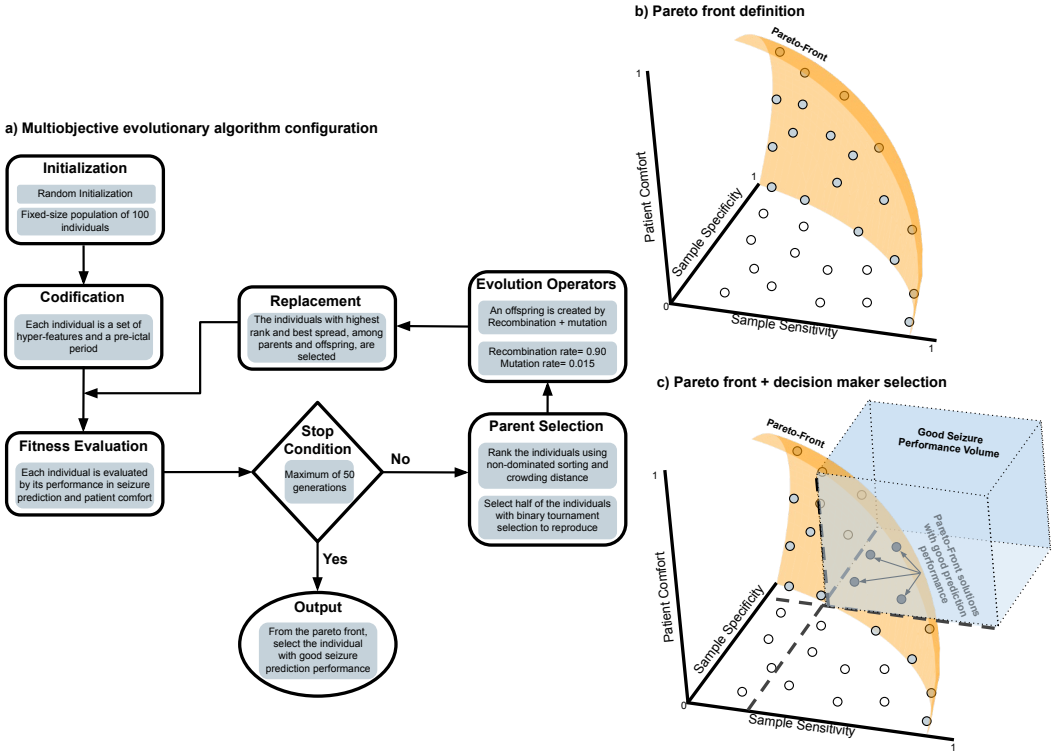


Figure 5.2: (a) The proposed MOEA configuration, (b) Pareto-front definition, and (c) the Decision Maker selection on the Pareto-front for selecting the individuals after the MOEA execution, where (b) and (c) reflect a problem formulated to simultaneously maximise all the objectives: two on seizure performance (sample sensitivity and sample specificity) and another on patient comfort.

tion) until generating 100 offspring. As variation operators, a recombination with a rate of 0.90 (90% of times, two parents produced an offspring) and a mutation with a rate of 0.015 (1.5% of offspring suffered a mutation) were used. This mutation rate was based on the number of genes per individual. Then, the individuals with the best rank and higher spread among parents and offspring were selected. These comprise the next generation of individuals. Evolution occurred over 50 generations. After the last generation, from the individuals with the best trade-off between the three fitness functions (Pareto-front or non-dominated individuals), only those with a seizure prediction performance higher than a determined threshold are saved.

Appendix K provides more information on the MOEA and a detailed explanation of the non-dominated sorting and crowding distance. Since a reduced computational time is needed to attain real-time applicability, this work was aimed for a fast convergence while achieving adequate solution diversity. Thus, that was the reason for opting for NSGA-II. Although this strategy may not usually produce an optimal solution, it is believed to approximate the global optimum within a reasonable amount of time [Deb et al., 2002, Cormen et al., 2001]. Each execution lasted approximately two hours on a machine equipped with an Intel Core i5-3230M 2.6GHz processor, 8GB of RAM, running on macOS Mojave 10.14.6 and using Python 3.7 on Spyder 4.0.1.

5.2.3.1 Encoding and variation operators

Figure 5.3.a) shows the idea behind genotype. A population is comprised of a group of individuals, where each is constituted by five hyper-features. Each hyper-feature is encoded with 13 genes: nine for the first-level feature (active feature group, active time feature, active frequency feature, active frequency band feature, statistical moment, Hjorth parameter, relative spectral power, wavelet coefficient energy, and spectral edge frequency), and a delay (minutes of a given feature before the preictal period), electrode, mathematical operator, and window length.

The mapping from genotype to phenotype consists in: i) finding the first-level feature that will be decoded to the phenotype for each hyper-feature, using the mentioned nine genes (as shown in Figure 5.3.b)); ii) constructing each hyper-feature by windowing the decoded first-level feature from the given electrode within the window length gene and, afterwards, by applying the respective mathematical operator; and iii) placing each hyper-feature chronologically in a timeline using the delay gene, according to the preictal period. This allows analysing a sequence of events with a given interval instead of the traditional analysis of feature alterations in that same interval. Since the preictal period is now included in the genotype, the MOEA also automatically searches for the optimal one. After constructing the five hyper-features and placing them chronologically, it is possible to evaluate a phenotype using the fitness function by performing a typical seizure prediction pipeline (sliding window analysis, classification, and regularisation). See Appendix L for an example of genotype-phenotype mapping.

Figure 5.3.c) shows all possible values for each gene, along with its neighbourhood. The latter is necessary to apply the variation operators, recombination, and mutation. The neighbourhoods were designed by considering the relationship between gene values (see Appendix N for more details on neighbourhood definition). Mutation, interpreted as a unitary step that will cause a random and unbiased change [Eiben and Smith, 2003], occurs in the following form for an individual (see Figure 5.3.d)): either one of the hyper-features or the preictal period gene is chosen randomly. When one of the hyper-features is chosen, one gene of that hyper-feature is then chosen randomly to mutate. If the preictal period gene is selected, its value will mutate. The remaining hyper-features and genes continue unaltered. Recombination is a stochastic operator that combines genetic information from two parents (individuals) into one or more offspring [Eiben and Smith, 2003].

After selecting two parents to reproduce, this operator performs the recombination of all paired hyper-features. Thus, hyper-feature pairing is the first step and, then, the recombination operator works at the hyper-feature gene level. Each offspring gene value is obtained by choosing a random one belonging to the shortest path between the correspondent two-parent gene values (see Appendix M for more details concerning evolution operators and an example for each).

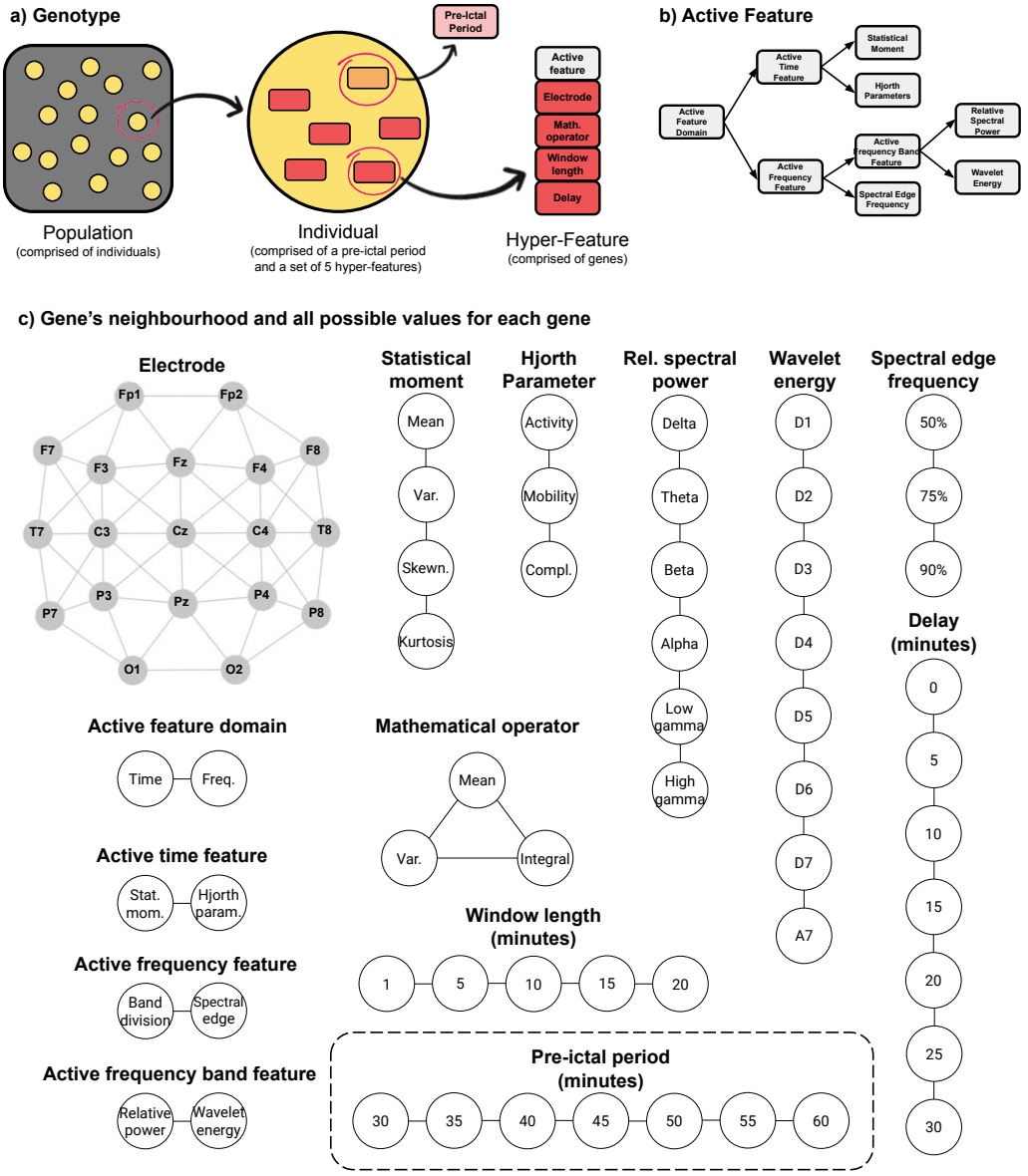


Figure 5.3: (a) Illustrated scheme of genotype. A population is comprised of a group of individuals, where each one is constituted by five hyper-features. Each hyper-feature is encoded with thirteen genes: nine for the first-level feature, delay, electrode, mathematical operator, and window length. (b) Active feature decoding, or in other words, how to decode the first-level feature from the genotype. (c) Gene's neighbourhood and all possible values for each gene, allowing the application of variation operators.

5.2.3.2 Fitness function

Figure 5.4 depicts the evaluation of each individual, which is performed iteratively (retraining the logistic regression classifier with new seizures) using the metrics typically used in seizure prediction. The seizures evaluated by the MOEA are referred to as validation seizures. After the MOEA, the same procedure is used to test new seizures, referred to as testing seizures (Figure 5.4.a). Thus, for each validation/testing seizure, the hyper-features and labels were extracted from the previous seizures

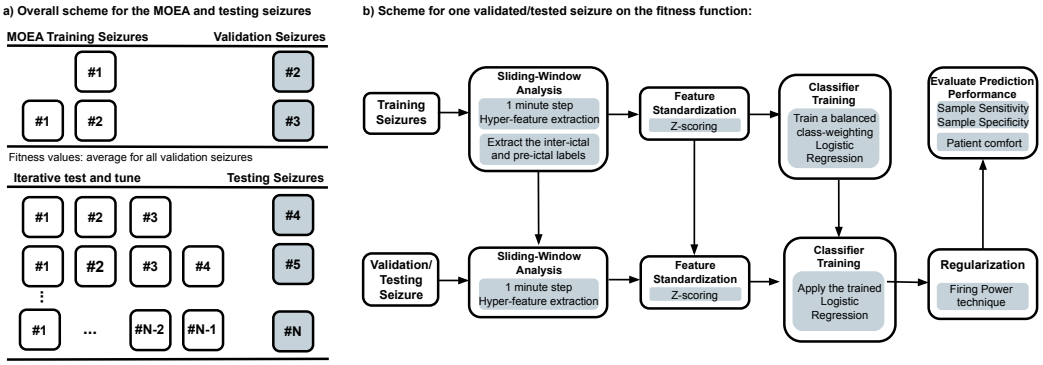


Figure 5.4: An illustrated scheme of the iterative retraining and validation in the MOEA input seizures. With the selected hyper-features, an iterative test and tune procedure is made (a). Scheme for one tested seizure on the fitness function (b).

through sliding-window analysis using a 1-minute step. Lastly, the feature set was standardised with z-scoring.

Then, the standardised features were used to train a classifier that balances samples according to a class weight inversely proportional to its frequency of occurrence. The chosen classifier was the logistic regression as it is an intrinsically interpretable model with low computational requirements [Molnar, 2019]. Next, a similar process was applied to the validation/testing seizure using all the parameters (mean and standard deviation for z-scoring) and models obtained from training.

Then, the Firing Power [Teixeira et al., 2012] regularisation technique was applied to smooth the classifier output and to make it more robust to noise. The latter works as a moving-average filter such that when a given threshold is exceeded, it triggers an alarm. It was decided not to include the Firing Power threshold in the MOEA genotype as it would be another parameter to tune and increase computational time. Therefore, a reasonable value of 0.7 was defined without any tuning.

The seizure prediction models were evaluated using three metrics. Two of them are based on seizure prediction performance: sample sensitivity SS_{sample} (Equation 2.1) and sample specificity SP_{sample} (Equation 2.2).

The third metric, Pcf, is related to patient comfort and concerns the number of different electrodes and their position on the scalp concerning lobes. It was assumed that more comfort to the patient (Pcf) could be provided by promoting solutions that do not require a large number of electrodes ($N_{electrodes}$) and that focus on a particular region of the brain (minimise the number of analysed lobes N_{lobes}). This last metric was computed according to the following Equation, where the term 1.25 is just a normalisation factor so that its range is $[0, 1]$:

$$Pcf = 1.25(1 - \frac{N_{electrodes} \times N_{Lobes}}{N_{electrodes} \times 5_{features}}). \tag{5.1}$$

This way, it is obtained a $Pcf = 1$ when only one electrode is used ($N_{electrodes} = 1$ and $N_{Lobes} = 1$). When five electrodes from different five lobes are used ($N_{electrodes} =$

5 and $N_{Lobes} = 5$), a $Pcf = 0$ is obtained. This aims to minimise the number of electrodes and the number of used lobes to maximise comfort. It is worth noting that there were other possibilities to handle spatial closeness, such as computing the total distance of each electrode to the remaining ones in the form of a graph. Nevertheless, this option was chosen to understand a possible relation concerning brain lobes.

After each MOEA execution has been completed, a Decision Maker (DM) was implemented to select which individuals from the set of Pareto-optimal solutions will be used in the testing phase. The DM selects individuals with a sufficiently high fitness score for the seizure prediction objectives (sample sensitivity and specificity) to use for testing. This restriction was implemented because even though the MOEA finds solutions corresponding to the analysis of a low number of electrodes/lobes, some of these Pareto-optimal solutions may present inadequate classification performance within the training set. Thus, a minimum fitness threshold of 0.9 was set for SS_{sample} and SP_{sample} metrics. When no solutions could be selected within a run, the threshold was decreased further to 0.8 to guarantee the selection of at least one individual. No minimum threshold was defined for patient comfort.

5.2.3.3 Training, testing, statistical validation, and performance comparison

In a real-life application, the MOEA would only be executed once, where it would select one individual from the Pareto-optimal set to predict new seizures. There is a need to understand how stochasticity affects performance. Thus, the MOEA was executed 30 times for each patient and then tested on unseen data. The testing phase was performed using the same pipeline for the fitness function (see Figure 5.4.b.) where it was also included a refractory period that follows an alarm with the duration of the preictal time. These periods were excluded from the False Prediction Rate per hour (FPR/h) metrics so that there was only considered the period during which false alarms can be triggered [Mormann et al., 2007].

The models' performance was assessed by calculating seizure sensitivity SS (Equation 2.3) and FPR/h (Equation 2.4). Then, a Surrogate analysis [Andrzejak et al., 2003] was performed to understand if the methodology was performing above the chance level. These metrics were also obtained for a control method inspired by a common machine learning prediction approaches, particularly in the work of [Cook et al., 2013]. Although the methods were built based on data from intracranial electrodes, it is the most relevant clinical trial that used an ML approach on a warning device. See Appendix P for more details on the control method.

The obtained patient models were statistically validated as follows: a method performs above the chance level when its seizure sensitivity is higher than the surrogate one, with a statistical significance of α=0.01 using a one-tailed $t - test$. It

was also verified if the set of validated patients had statistical significance in the same way as [Alvarado-Rojas et al., 2014]. Considering a statistical significance of α=0.05, the probability of observing for at least i of I (patient-models) executions that outperformed the surrogate predictor was given by:

$$P_{binom}(i, I, \alpha) = \sum_{j=i}^{I} \binom{I}{j} \alpha^j (1-\alpha)^{(I-j)}. \tag{5.2}$$

5.2.3.4 Phenotype study

A phenotype study was performed to analyse the overall independent influence of each gene value on the population. Thus, a binary value (0 or 1) was assigned for each gene value that corresponds to its presence in a hyper-feature. The presence binary value was calculated for all hyper-features from selected Pareto-optimal individuals (see Appendix O for more details and mathematical formulation).

However, these metrics cannot provide information on interaction, i.e., information regarding which features always appear in the presence of others. Therefore, it was also implemented the *apriori* algorithm [Borgelt and Kruse, 2002] that aims at finding frequent patterns in the obtained phenotypes (association learning). By using association rules, the goal was to find subsets of gene values frequently appearing together and, therefore, have a high probability of describing seizure generation processes.

5.3 Results

This section is divided into two subsections. The first presents the results for the MOEA and control method, and the second, the phenotype study.

5.3.1 MOEA and control method performance

Figure 5.5 depicts the results for the testing seizures, both for the MOEA and the control method, along with patient stratification. Colour represents seizure sensitivity (0-1), while the diamond-shaped marker represents the patient models that outperformed the surrogate predictor, or in other words, performed above chance.

The MOEA performed above chance for 30 patient models (32%), while the control method validated 33 patients (35%). By inspecting Figure 5.5 and the test results in full detail (see Appendix Q), it is possible to see that the MOEA obtained lower sensitivities and lower FPR/h values when compared to the control method. Although the MOEA obtained six models that were statistically validated while presenting an adequate FPR/h (<0.15 [Winterhalder et al., 2003]), high sensitivity is missing for claiming its use in real-life. However, the control method presented models for eight patients (202, 3300, 6000, 8902, 21902, 26102, 1310803, 1322803)

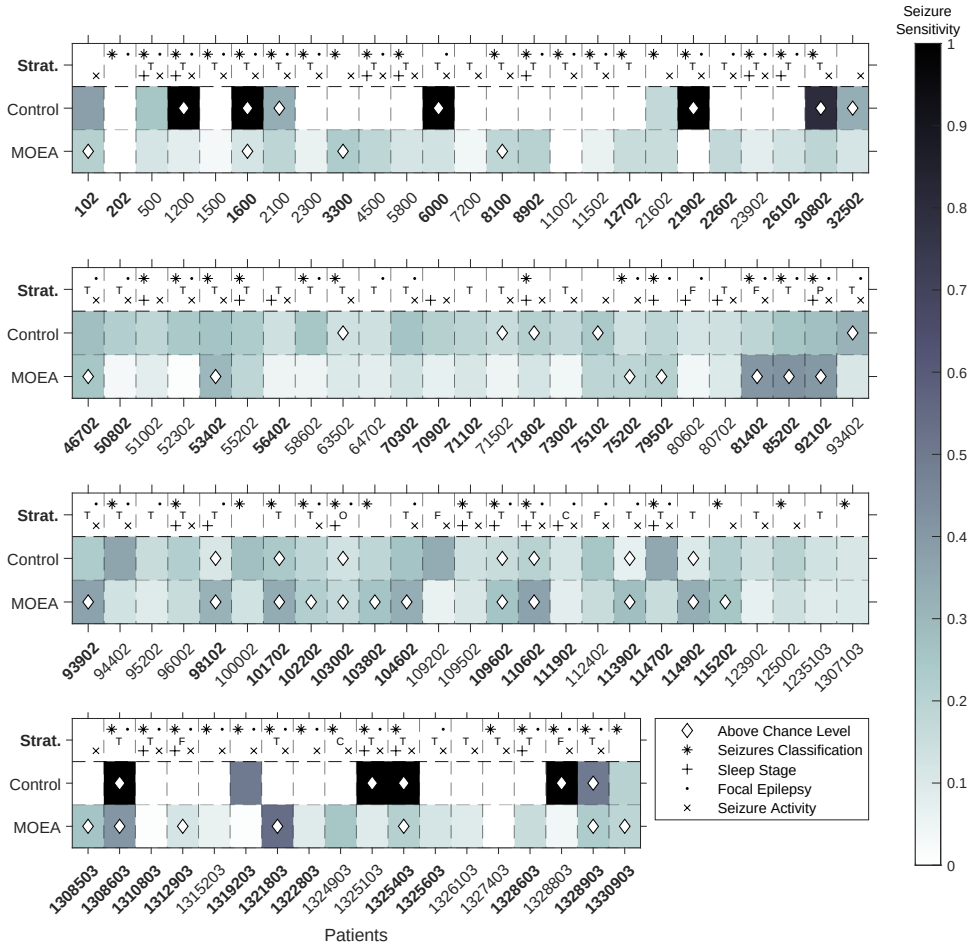


Figure 5.5: The test performance of all patient models, for the MOEA and the control method. Colour represents the seizure sensitivity, while the diamond shape means that the models performed above the chance level. On the top line (Strat.), patient stratification is presented concerning seizure classification (*, Focal Onset Aware (FOA) or Focal Onset Impaired Aawareness (FOIA)/focal to bilateral tonic-clonic seizures), sleep stage at seizure onset (+, awake/sleep), type of epilepsy (., focal/generalised) and annotated activity pattern (x, rhythmic/non-rhythmic), and seizure focus that is lobe-specific (temporal (T), frontal (F), central (C), parietal (P), occipital (O)). Patients in bold achieved performance above-chance for either the MOEA or the control method.

with maximum sensitivity and adequate FPR/h. Nevertheless, it is worth noting that seven of these patients only had one seizure for testing, while patient 8902 had two.

It is important to mention that these methods might be overfitted (overestimated) to the training seizures, as the training results are considerably higher than testing. While the average fitness in the validation seizures (Appendix Q) was 0.97±0.02 for sample sensitivity and 0.96±0.02 for sample specificity, for testing, the MOEA obtained an average seizure sensitivity of 0.16±0.11 and an average FPR/h of 0.21±0.08. Concerning patient comfort (Appendix Q), it is worth noting that the ideal scenario of having only one electrode was not achieved in any patient model. In fact, the patient comfort objective ranged from 0.50±0.19 (patient 111902) to

Table 5.1: Test results for the overall set of patients (sensitivity, FPR/h, and ratio of patients performing above chance, both for the MOEA and the control method), and for stratified sets of patients. Stratification concerned patients with the following criteria: i) focal epilepsy patients (seizure focus in only one hemisphere), ii) generalised epilepsy patients (seizure focus in more than one hemisphere), iii) temporal lobe epilepsy patients (seizure focus only in the temporal lobe), iv) patients whose seizure focus is not lobe-specific, v) patients suffering only from FOA/FOIA seizures, vi) patients whose surgery was offered (localised focus), vii) patients whose surgery was not offered (not localised focus), viii) patients that only experienced seizures while awake., and ix) patients with pre-seizure activity annotated as rhythmic by clinicians.

		MOEA			Control		
Stratification	**Number of Patients**	**SS**	**FPR/h**	**Patients performing above chance (0-1)**	**SS**	**FPR/h**	**Patients performing above chance (0-1)**
Focal Epilepsy	55	0.15±0.12	0.21±0.08	0.33	0.31±0.39	0.67±0.92	0.35
Generalised Epilepsy	33	0.17±0.11	0.21±0.09	0.33	0.35±0.37	2.18±7.13	0.36
Temporal Lobe Epilepsy	61	0.16±0.12	0.22±0.09	0.30	0.32±0.40	1.54±5.30	0.33
Non-specific lobe	**21**	**0.15±0.08**	**0.21±0.06**	**0.38**	**0.38±0.37**	**1.01±2.31**	**0.48**
Only FOA or FOIA seizures	**59**	**0.17±0.12**	**0.22±0.07**	**0.39**	**0.28±0.38**	**0.82±1.57**	**0.31**
Localised focus	49	0.15±0.10	0.22±0.08	0.31	0.28±0.36	0.66±0.91	0.29
No localised focus	**44**	**0.15±0.12**	**0.20±0.08**	**0.30**	**0.37±0.39**	**2.18±6.33**	**0.45**
Awake-only onset seizures	29	0.16±0.09	0.21±0.07	0.31	0.33±0.38	1.05±2.04	0.38
Rhythmic activity only	59	0.15±0.11	0.22±0.09	0.29	0.32±0.36	1.54±5.45	0.37
Overall	**93**	**0.16±0.11**	**0.21±0.08**	**0.32**	**0.32±0.38**	**1.31±4.43**	**0.35**

0.86±0.11 (patient 10962). Thus, the MOEA opted for three to four electrodes to capture pre-seizure brain dynamics.

Table 5.1 shows test results for the set of patients and patient stratification. This stratification was based on seizure classification FOA or FOIA only), seizure activity (only rhythmic), vigilance state at seizure onset (awake only), seizure focus in one hemisphere/more than one hemisphere (focal/generalised epilepsy), temporal lobe epilepsy, seizure focus not restricted to a single lobe, and concerning surgery decision (localised/not localised). Due to the small number of patients, other stratified groups were not considered. For example, there were only seven patients available with frontal lobe epilepsy. In the stratification regarding seizure classification, activity, and vigilance state at seizure onset, a patient was selected if a given criterion was met both in the training and testing seizures.

Seizure classification and the patients with a seizure focus in non-specific lobes were the only criteria that considerably improved the percentage of validated patient models (39% and 38%, respectively) for the MOEA. Moreover, in the FOA or FOIA seizures stratified groups, the ratio of validated patient models was higher on the MOEA methodology than on the control.

5.3.2 Phenotype study

The patients' phenotype was analysed to demonstrate the MOEA potential to unravel pre-seizure knowledge. This subsection presents a phenotype study for all patients, which explores gene importance (Figures 5.6 and 5.7). The results indicate that most solutions were obtained from three electrodes across two lobes and localised in two regions (three regions were considered: left and right hemispheres and the central part).

An additional analysis was also performed (see Figure 5.8): the inspection of the

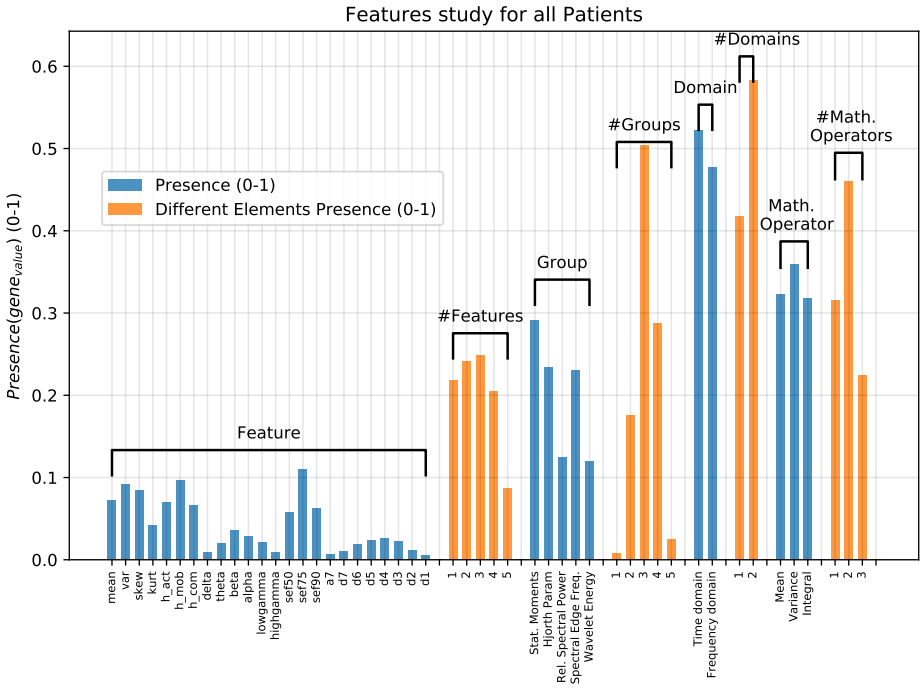


Figure 5.6: Phenotype feature study for all patients. The presence of gene values is presented in blue. The simultaneous presence of different gene values is presented in orange.

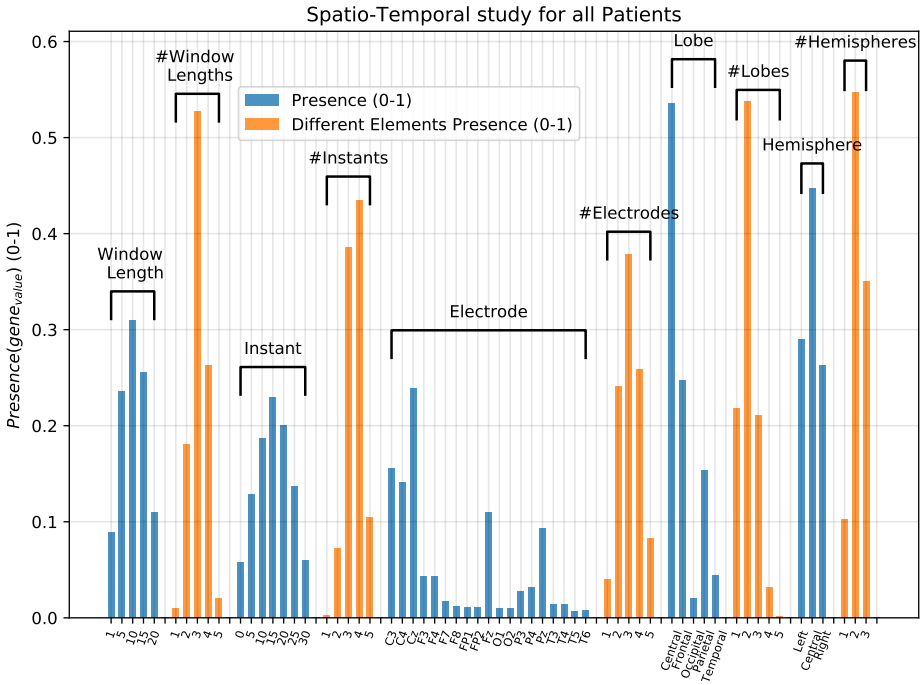


Figure 5.7: Phenotype temporal and spatial study for all patients. The presence of temporal (window length and time instant) and spatial (electrode, lobe, and hemisphere) gene values are presented in blue. The simultaneous presence of different gene values is presented in orange.

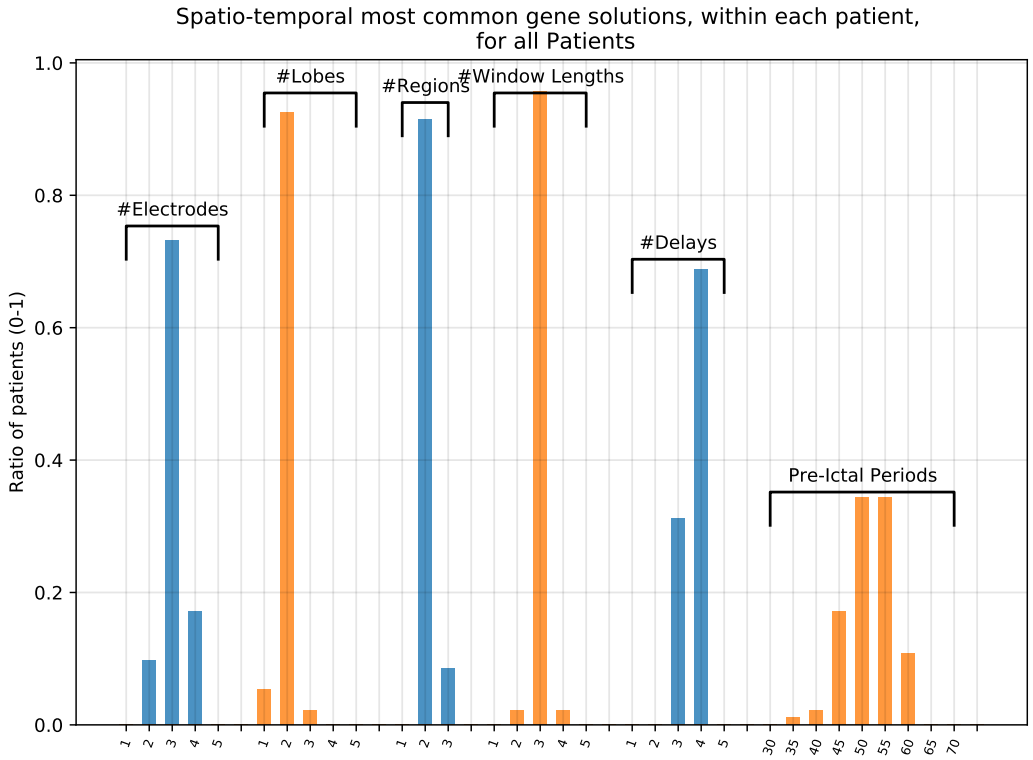
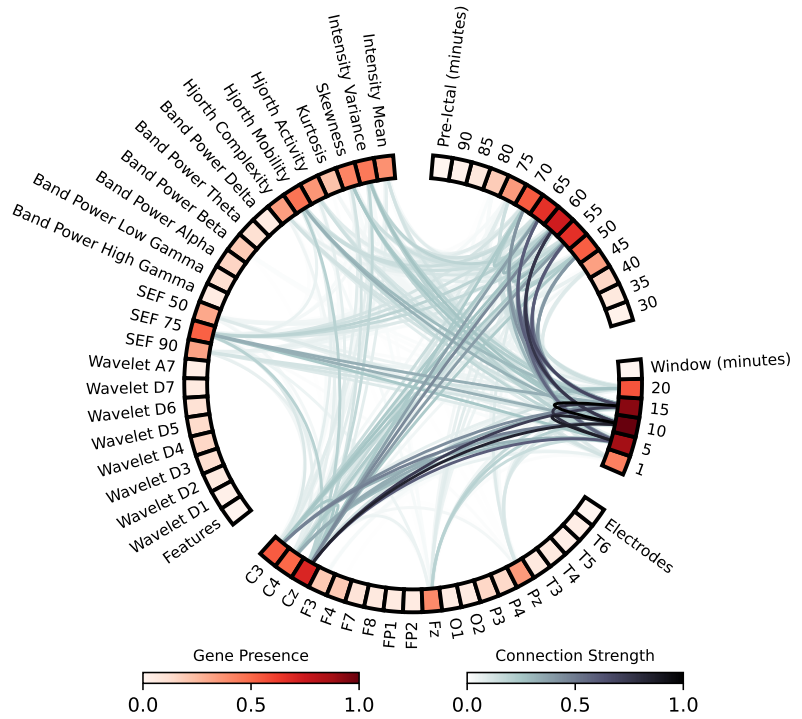


Figure 5.8: The most common gene solutions, for each patient, for all patients concerning the number of simultaneously used different electrodes, lobes, regions, window lengths, and delays. The most common preictal period for each patient is also presented.

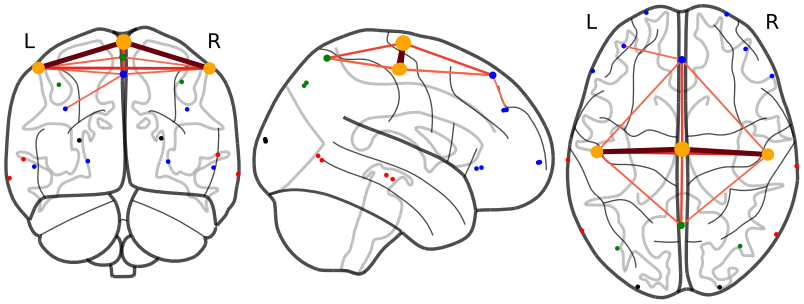
most common solutions within each patient, particularly the obtained preictal period and the number of different electrodes, lobes, regions, window lengths, and delays. With this analysis, it is possible to confirm that most patients had a set of three electrodes, two lobes, and two regions as the most common solution. Moreover, no patient had the same window length for all features and used at least three delays. The most common preictal periods were 50 and 55 minutes, despite some presenting a 60- or a 45-minute one.

Additionally, gene interaction was studied by showing which gene values appear along with other gene values and how to present them intuitively. Figure 5.9 (a) represents the gene presence (red colourmap) and gene interaction (gray colourmap), and (b) represents brain connectivity. Gene interaction was calculated using the *apriori* algorithm [Borgelt and Kruse, 2002]: first, frequent associations were found between gene values, and then the association lift measure [Borgelt and Kruse, 2002] was used to map interaction strength. Lift ratios larger than 1.00 were used, which means that the two association items are more likely to appear together than separated. The larger a lift ratio is, the more significant is a given association rule. Brain connectivity was computed using association rules between electrodes and association lifts higher than 1.0.

The most important features were Spectral Edge Frequency (SEF) 75, Hjorth



(a)



(b)

Figure 5.9: Gene interaction study for all patients. (a) Gene presence (red colourmap) and gene interaction (grey colourmap) in a chord diagram plotted using MNE Python library [Gramfort et al., 2013]. Gene presence is the ratio of times a given gene is present. Gene interaction is given by summing the association lift measures calculated with the *apriori* algorithm and then normalised to a 0-1 scale. (b) Brain connectivity was plotted using the *Nilearn* Python library, and the 10-20 system electrodes were converted to the MNI coordinates [Okamoto et al., 2004]. Node size corresponds to the electrode presence, while edge colour and thickness provide connection strength (association lift). Node colour represents brain lobes (blue: frontal, orange: central, green: parietal, black: occipital, and red: temporal). *Apriori* algorithm parameters: minimum support of 0.07, minimum confidence of 0.10, and minimum lift of 1.00.

Mobility/Variance, and skewness. These features have the highest gene presence and present the highest number of strong interactions. The same gene interaction study was conducted for the electrodes (with a prevalence of C3, C4, CZ, PZ, and Fz) and preictal periods (prevalence of 45 to 75-minute intervals). Concerning window length, all windows were present in phenotype, with a higher frequency observed for the 10- and 15-minute windows. Gene interactions between feature-to-feature,

electrode-to-electrode, and window-to-window genes were also found. The most robust relations were observed for the window-to-window gene interactions.

These findings demonstrate the importance of simultaneously analysing different time windows. It is possible to obtain a similar conclusion by inspecting Figure 5.7, where the MOEA chose different delays about 99% of the times and 99% of the times concerning different window lengths. Electrode-to-electrode interactions were also observed, which can be interpreted as a manifestation of brain connectivity. Some connections arise across the central, frontal, and parietal lobes. The whole-brain state's importance demonstrates the difficulty of finding a good set of predictive features while providing maximum patient comfort (minimising the number of electrodes and restricting electrode placement region). In the case of this work, in addition to having patient comfort as an objective of the MOEA, it was considered pertinent to inspect brain activity across different lobes and hemispheres.

5.4 Discussion

Despite the control method being marginally better than the proposed MOEA in terms of validated patients, it is believed that the MOEA has potential and brings added value. It is interesting to verify that only nine patients had both methodologies (MOEA and control) performing above chance. It was possible to find statistically validated models for 54 patients (58%). Of these 54 patients, 21 (≈38) were only validated by the MOEA. These results suggest the need to vary the selected features and classification models and use significantly different methodologies in a complementary way to capture patient-specific pre-seizure patterns. Moreover, from the 33 validated patients by the control method, six (1328603, 1325603, 1308503, 109602, 98102, 71802) presented an FPR/h>1.50, and ten patients presented an FPR/h>1.00 (1328903, 1328603, 1325603, 1308503, 109602, 98102, 73002, 71802, 71102, 50802). In the MOEA, no patient presented a mean FPR/h higher than 0.37, although the sensitivity values were significantly lower. Lastly, the obtained model from the control methodology is a random forest with several features, ranging from 3 to 20 features. In fact, the average number of used features was 9.25±6.26. The MOEA only uses five features for computational reasons, limiting the number of electrodes to five.

Although the evaluation metrics of the MOEA evaluation metrics differ from the testing ones, high sample performance leads to high seizure sensitivity and low FPR/h. This evidence can be explained due to confounding factors, as the used database concerns patients under presurgical monitoring submitted to medication withdrawal to study the epileptic focus. When exploring the obtained results, there is an interesting relation between the number of tested seizures and validated patient models. While there was computed a negative correlation coefficient ($\rho = -0.15$) when analysing the control method, an almost null coefficient ($\rho = -0.01$) for the

MOEA was obtained. These results suggest that retraining the logistic regression after a new seizure occurs might improve seizure performance. This strategy intrinsically assumes that the MOEA selected features can deal with new concept drifts, and that is why the fitness function is based on iterative training. It is believed these results would be more significant by iteratively running the MOEA to update the set of selected features instead of just training the logistic regression with upcoming tested seizures. By re-selecting new features, it would better handle concept drifts. This strategy was not performed due to computational costs.

Nevertheless, performing MOEA update for each new upcoming seizure can be considered for real-time applications, as running one execution of the MOEA takes approximately two hours, more than twice shorter than the analysed inter-seizure independence interval (4h30min). This methodology was only tested once, and there was no attempt to improve results. It was also decided to keep a 10-minute intervention time [Winterhalder et al., 2003] to provide the patient sufficient time to prepare for a seizure. This period is also compatible with the intake of diazepam rectal gel (the only FDA drug approved for seizure cluster and which might be tested for prevention) that needs 5-10 minutes to take effect [Dreifuss et al., 1998].

There are some studies presenting seizure prediction models developed that use EPILEPSIAE [Bandarabadi et al., 2015b, Teixeira et al., 2014b, Alvarado-Rojas et al., 2014, Rasekhi et al., 2013]. Some works [Bandarabadi et al., 2015b, Teixeira et al., 2014b,Rasekhi et al., 2013] cannot be directly compared to the proposed methodology, as they developed several models and presented the best ones based on testing results, which do not correspond to real-life performance. The study from [Alvarado-Rojas et al., 2014] is comparable with this methodology due to the following: i) model selection was not based on testing performance but on training data only; ii) it contains 53 patients (in total 531 days and 558 clinical seizures); and iii) they used a threshold classifier, which is intuitive, despite the use of features that may be difficult to understand from a clinical point of view. Their results outperformed in seizure sensitivity (≈ 0.66), but the MOEA was better in terms of FPR/h (≈ 0.33) and performance above chance level for a higher ratio of patients ($7/53 \approx 0.13$). The proposed methodology is also more intuitive: each model uses a binary decision based on a threshold (logistic regression) and comprises five widely-used and intuitive features. [Direito et al., 2017] also allows a good comparison, as it reports the analysis of the largest group of patients from EPILEPSIAE (218 patients). Their work also outperformed in sensitivity (0.39), but statistical validation was achieved for about 11% of the patients. Nevertheless, it is worth noting that they used a random predictor [Schelter et al., 2006] to perform statistical validation instead of the surrogate analysis.

It is worth noting that the obtained performance is still far from being ideal for a real-time application, despite the ratio of validated patient-specific prediction models remaining similar to the first published paper from this methodology [Schulze-

Bonhage et al., 2010]. It is worth noting that three patients (16202, 60002, 98202) of [Pinto et al., 2021a] were discarded in [Pinto et al., 2022], as the correspondent recordings presented significant gaps of missing data (over one hour). In terms of sensitivity, the results were also not satisfactory. Concerning FPR/h, the new results were better. Due to this, these models might not fully capture the complexity of the seizure generation processes. Results also changed within the same patient when comparing both works. Patient 55202 is one example: it presented significant EA results, but not with the MOEA. The first 60% of seizures were used in [Pinto et al., 2021a] to train the algorithm, and the last 40% were used for testing. In [Pinto et al., 2022], since it was desired to have more testing seizures, the first three seizures were used to train the algorithm, and the remaining ones were tested. Patient 55202 has eight seizures. Therefore, the first EA was trained with five seizures and tested with three seizures, while the MOEA was trained with three seizures and tested with five. Performance may increase from having more seizures as training data.

The use of Deep Learning (DL) may theoretically yield improved results by enabling more complex and hence less intrinsically interpretable models, however, at the cost of losing clinical interpretation [Freestone et al., 2017]. Furthermore, the statistical validation method limits these study comparisons with this work. All of the studies mentioned above use the random predictor [Schelter et al., 2006] while this study used a surrogate predictor. The latter is flexible, and it adapts better to the data. Moreover, the choice of the surrogate predictor was bounded by the considerable number of models obtained for each patient and the small number of tested seizures per patient. Having few seizures to test the models can significantly discourage using a random predictor [Mormann et al., 2007, Andrzejak et al., 2003].

As previously mentioned, efforts were made toward building models to be applied in a real-time scenario, using a 10-minute intervention time and extracranial recordings. This work focused on scalp EEG as it was desired to understand the importance of the whole brain state and not only the seizure onset zone. This might explain that this study of many different types of epilepsy has the same outcome as the earlier study of temporal lobe epilepsy only. It is worth stressing that iteratively re-selecting the features by executing the MOEA periodically would consider a dynamic epileptic network rather than a static one, which may produce more insights into the brain dynamics [Kuhlmann et al., 2018b].

5.4.1 Added value

This work aims to comply with current legislation, specifically with the 2018 General Data Protection Regulation (GDPR) for European citizens and the European economic space. One of the proposed seizure prediction model's main advantages is exploring the obtained phenotype to find patients' preictal patterns.

Including the patient comfort metric has theoretically allowed obtaining more

comfort EEG configurations for most patients. Since it was obtained a set of three electrodes for 68 patients (73%) and a set that uses two lobes for 86 patients (92%), it is possible to claim that patient comfort was relevant within the MOEA. Nevertheless, the used strategy for patient comfort might not be the best. The optimal way to conclude about the comfort relevance would be the following: i) execute the MOEA for all patients without the comfort objective, ii) by performing the phenotype same study on the number of the electrodes and lobes, and iii) by comparing the obtained output. It is worth noting that not all MOEA solutions present the same set of electrodes within each patient. Thus, one can choose one configuration that was not the most common within the obtained solutions as long as it provides more comfort and maintains its performance levels. To better understand the relevance of the number of chosen electrodes and lobes, one can visualise a histogram for all solutions, for all patients, on this paper's GitHub page. Five examples (patients 1200, 12702, 55202, 81402, and 1319203) are also presented in Appendix S. An overall study can also be seen in Figure 5.8, developed using the mode operator (most frequent value) for each patient. Five patients were chosen as these cases represent all found case scenarios concerning the electrode/lobe topic. Additionally, when analysing the number of occupied regions (left hemisphere, central part, right hemisphere), the majority of occupied regions is two. This limitation must be addressed in the future by including a factor that minimises the number of occupied regions in the patient comfort objective.

The impact of patient comfort was analysed on performance. For each patient, two scatter plots were made to access the patient comfort impact: i) scatterplot(comfort, sensitivity) and ii) scatterplot(comfort, FPR/h). The Pearson correlation coefficient between comfort and sensitivity and between comfort and FPR/h was also computed to assess this impact. Null correlation values were obtained for sensitivity and FPR/h when analysing the overall correlation between all patients. However, it was possible to find several case scenarios by analysing each patient separately. For 33% of patients, overall performance increased with more comfortable electrode configurations. For 32% of patients, overall performance was maintained, and for 35% of patients, overall performance decreased with more comfortable configurations. It was considered a sensitivity increase/decrease with comfort when the correlation between sensitivity and comfort was higher/lower than 0.10/-0.10. When the absolute correlation value was lower than 0.10, performance was considered to be maintained despite the used electrode configuration. A similar rationale was applied to FPR/h: it was considered a significative sensitivity increase/decrease with comfort when the correlation between sensitivity and comfort was higher/lower than -0.10/0.10.

A decrease in performance was expected, and maintenance could be expected. Nevertheless, an increase in performance was not expected. This increase could be due to the overfitting of configurations using more electrodes. By using fewer

electrodes, model generalisation may be more easily achieved. Thus, it may be possible to conclude that, for some patients, the most comfortable configurations may be used as performance is not negatively affected. The analysis for each patient can be found on this paper's Github. Appendix T shows some examples (patients 1500, 58602, and 21602).

5.4.2 Study limitations

Due to computational speed, it was only used the last four hours of data before each training seizure. This is the most significant limitation of this methodology, as it is strongly advised [Mormann et al., 2007] to use all available data from training seizures. Using all available data would increase interictal data representativity and identify and deal with reoccurring confounding variables. Also, the remaining parameters, such as the 1-minute step and number of used features, were reasonably chosen based on the computation time, without any tuning on the test results. Even though the results may be low in terms of sensitivity and FPR/h, they concern presurgical monitoring of long-term data without preprocessing, except filtering.

It is also essential to discuss the obtained preictal periods. All patients presented a similar preictal period of around 50 minutes. All solutions' preictal periods were analysed from all patients. Appendix R presents some examples for patients 21602, 21902, 30802, and 32502, which had, as mean preictal period, the following values, respectively: 49.91±9.55, 51.40±9.55, 51.73±8.31, and 53.14±8.70 minutes. The most frequent preictal period (the mode) for each patient is also presented. The MOEA can find solutions for many possible preictal periods in all patients. However, when analysing the most frequent one (the mode) for each patient, it can range from 35 to 60 minutes (see Figure 5.8). An overall can also be seen in Figure 5.8, developed using the mode operator (most frequent value) for each patient.

It may be desirable to have a specific preictal period for each patient. However, choosing an MOEA solution may also be helpful based on the patient preferences concerning a trade-off between performance and the preictal period. More extended preictal periods may bring better performances but may also cause considerable stress. It is believed these findings are a built-in bias. One hypothesis may be that the preictal period changes the number of samples belonging to each class (preictal, interictal) in problems with high-class imbalance (as is the case of seizure prediction), and thus sensitivity and specificity might be significantly affected. As sensitivity and specificity are metrics used in the MOEA, it is very likely the existence of a built-in bias during the search procedure of the algorithm that drives the preictal period towards values that optimise different trade-offs between sensitivity and specificity through class balancing.

Also, 31 of the used patients had only one seizure for testing. This study tried to make a good compromise between data quality and the number of patients, which

resulted in 93 patients. Due to data constraints, it is difficult to perform a seizure prediction study, particularly patient-specific, with such a high number of patients. If the minimum number of seizures was raised to five, 31 patients would have been discarded, representing a third of all used patients. On average, there were tested 2.57 seizures per patient.

Lastly, using five electrodes may result in several patients' severe undersampling of the underlying cortex. However, this work is not the first to tackle patient comfort by limiting the number of electrodes. In contrast, other authors have used three to six electrodes to experiment with different configurations [Direito et al., 2017, Bandarabadi et al., 2015b, Teixeira et al., 2014b].

5.4.3 Final reflections

This work contributed to epilepsy seizure prediction by providing a complete pipeline for patient-specific prediction while addressing concerns regarding patient comfort through electrode placement on the scalp. More importantly, it shows that there is potential to develop different strategies for improving ML model communication to clinicians. Although this study used an intrinsically interpretable model (logistic regression), authors can choose any black-box model with this methodology as the phenotype study allows to retrieve knowledge from any selected model.

This study includes 93 patients with several types of focal epilepsies and generalised epilepsy, but the data concerns presurgical monitoring conditions. Due to time constraints in the clinic, patients suffer medication reduction and sleep deprivation, which may induce more seizures that may not be representative of daily-life events [Freestone et al., 2017]. To truly access seizure prediction performance, it is necessary to replicate this study on ultra long-term recordings collected during the patients' daily life, such as the collected data from the Neurovista prediction challenge [Kuhlmann et al., 2018a] and in a proper field test. These two steps would answer how far this methodology may perform and how to apply it to a medical device at home. Concerning patient outcomes, it is worth noting the awareness of these limitations.

The access to more computational resources may improve this methodology, for example, by enabling the increase of the number of features and/or the test of more complex classifiers. However, one must be careful with publication bias by reporting higher performances, as these may result from a test set optimisation. Also, these improvements will increase the execution time of the MOEA, which should be kept to a reasonable period. A higher runtime may imply the redefinition of some parameters, such as the number of generations and population size. Changing these aspects will not influence the application of the association rules, which can retrieve information about gene importance and interactions. It is believed that this methodology may significantly benefit from information regarding concept drifts,

such as medication intake and circadian cycles [Kuhlmann et al., 2018b].

Chapter 6

Explaining Machine Learning prediction models

This chapter concerns how to explain Machine Learning (ML) models for Electroencephalogram (EEG) seizure prediction and how to evaluate the developed explanations with clinicians and data scientists, from Deep Learning (DL) models to intrinsically interpretable ones. This chapter resulted in five lessons to help researchers explain their models to clinicians. Section 6.1 presents a study context. Section 6.2 details the used methods and materials. Section 6.3 provides the obtained lessons and how to apply them. Lastly, section 6.4 discusses the obtained findings and limitations, and provides final reflections.

The content of this chapter is under preparation to be submitted to a scientific journal. The code for this study is available at https://github.com/MauroSilvaPinto.

6.1 Study context

Despite explanations having a sociological component [Molnar, 2019, Doshi-Velez and Kim, 2017], current ML studies on EEG explainability lacked rigorous validation as they did not formally evaluate how the created explanations helped specialists complete their tasks. Explanations are an exchange of beliefs that help to answer a "why" question when one can no longer keep asking why [Miller, 2019, Molnar, 2019, Gilpin et al., 2018, Miller et al., 2017].

In this chapter, to properly evaluate and understand what may be the most critical model explanations for EEG seizure prediction, three ML solutions with different levels of complexity were developed and quasi-prospectively evaluated. Then, patients were selected to construct different explanations for their model decisions. Explanations were presented to ML experts (data scientists working on clinical problems) and clinicians (neurologists and EEG technicians working in an epilepsy refractory centre).

Results showed that current explanation methods are insufficient to understand

brain dynamics, even when using intrinsically interpretable models with few widely used features. As clinicians cannot detect pre-seizure patterns several minutes before seizures, the goal of explainability is not merely to explain model decisions but to help improve the system. Explainability allows researchers to design hypotheses regarding physiology and model behaviour that help develop a complete problem formulation. Even when pre-seizure underlying mechanisms are unknown, one obtains more trust in the models when testing the developed hypotheses, and they still stand.

6.2 Materials and methods

The used methodology comprises three main steps: i) developing ML methodologies, ii) developing explanations, and iii) evaluating explanations. Summarily, three different pipelines were developed with different degrees of transparency: a class-weighted logistic regression model, a voting ensemble of 15 Support Vector Machiness (SVMs), and a voting ensemble of three Convolutional Neural Networkss (CNNs). Then, patients with high and low performances were selected to develop explanations about their models' functioning and decisions. The interviews were performed by showing the developed explanations and asking open-ended questions about them to scientists working in healthcare and clinicians (neurologists and EEG technicians) working in an epilepsy refractory centre. The interviews were analysed using Grounded Theory (GT), which was vital to understand model explanations and their significance in EEG seizure prediction research. The findings were summarised in five lessons by interpreting the emerged themes and ideas.

6.2.1 Dataset

From the EPILEPSIAE database [Klatt et al., 2012, Ihle et al., 2012], 40 patients with Drug-Resistant Epilepsy (DRE) were selected (23 males and 17 females, aged 39.42±15.87 years) from the Universitätsklinikum Freiburg in Germany. In total, this dataset contains 224 seizures (120 for training and 104 for testing), 3254 hours of training data ($\approx$ 4.52 months), and 1402 hours of testing data ($\approx$ 1.95 months). The patient selection criteria were: i) patients with temporal lobe epilepsy, as it concerns the most common type of focal epilepsy [Rubboli and Gardella, 2019]; ii) a minimum of four lead seizures separated by periods of at least 4h30; and iii) EEG scalp recorded with a sampling frequency of 256 Hz. Electrodes were placed according to the 10-20 system. The data were collected while patients were in the clinic for presurgical monitoring. The Ethical Committee approved the use of this data for research purposes of the three hospitals involved in the database development (Ethik-Kommission der Albert-Ludwigs-Universität, Freiburg; Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine

de la santé, Pitié- Salpêtrière University Hospital; and Ethics Committee of the Coimbra University Hospital).

All methods were performed following the relevant guidelines and regulations. Informed written patient consent from all subjects and/or their legal guardian(s) was obtained to participate. Informed consent was obtained from all subjects and/or their legal guardian(s) for online open-access publication. See Appendix U for detailed information on the selected patients.

6.2.2 Developing ML methodologies

The three methodologies use the rationale from Figure 6.1. A CNN developed with EPILEPSIAE database [Lopes et al., 2021] preprocessed the raw EEG recordings. Then, univariate linear features were extracted from time and frequency domains within five-second sliding windows without overlap. For the frequency domain, the following features were extracted: relative spectral power bands (delta (0.5-4Hz), theta (4-8Hz), beta (8-13Hz), alpha (13-30Hz), gamma band 1 (30-47Hz), gamma band 2 (53-75Hz), gamma band 3 (75-97Hz), and gamma band 4 (103-128 Hz)), the ratio between these bands, spectral edge frequency and power at 50%. For the time domain, the following features were extracted: the four statistical moments (mean, variance, skewness, kurtosis), Hjorth parameters (activity, mobility, complexity), and decorrelation time. The energy of the wavelet coefficients (from D1 to D8, using the db4 mother wavelet) was also extracted. More details on the extracted features and expected added value can be found in Appendix B.

All methodologies were patient-specific, where each patient's first three chronological seizures were selected for training (preictal period estimation, standardisation, class balancing, feature selection, and model training) and the remaining ones for testing (as in Figure 6.1). A grid search was used to obtain the preictal period (30 to 60 minutes in steps of 10) [Assi et al., 2017,Mormann et al., 2007]. For testing seizures, it was also applied the Firing Power [Teixeira et al., 2012] to smooth model predictions over time, where an alarm threshold of 0.7 was reasonably defined [Pinto et al., 2022,Pinto et al., 2021a]. After each alarm, a refractory period of equal length to the preictal period was applied. Performance was evaluated [Winterhalder et al., 2003] by calculating seizure sensitivity (SS, Equation 2.3) and False Prediction Rate per hour (FPR/h) (Equation 2.4) and performing a surrogate analysis [Andrzejak et al., 2003]. A seizure prediction horizon [Winterhalder et al., 2003] of 10 minutes was used to allow patient intervention.

The most straightforward methodology used a Logistic Regression, where the classes' weight was balanced in inverse proportion to their frequency (the preictal class has more weight as it contains fewer samples than the interictal one). The F-test [Kramer, 2016] was used for feature selection. This methodology was deterministic. Two stochastic methodologies were also developed: one using a voting

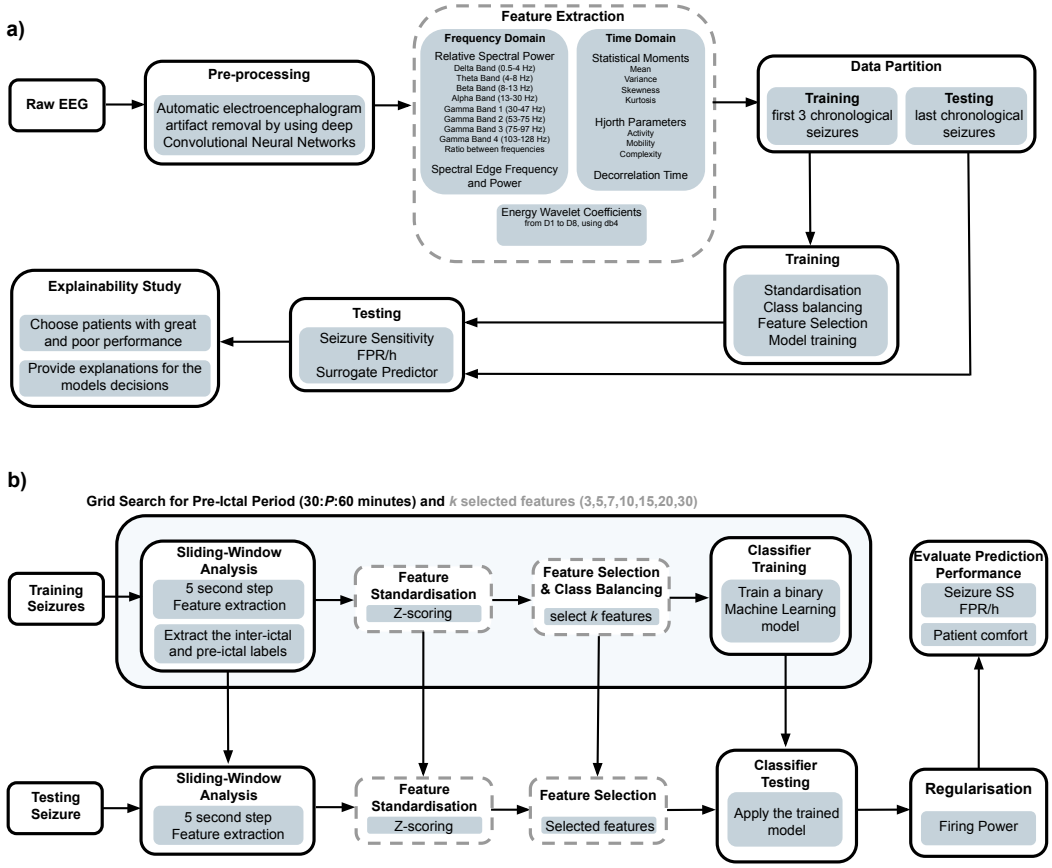


Figure 6.1: A general overview of this study's work for all three developed methodologies. Part a) presents the general strategy for the entire study. Part b) details the training and testing phases of the ML pipeline. Steps represented by grey dashed bordered boxes and grey text (feature extraction, selection, and standardisation) are performed only for the methodologies that include the SVM and the logistic regression models.

system of 15 SVMs and another a voting system of three CNNs. These two were balanced by randomly selecting samples from equally spaced segments over the signal to get a representative set of the interictal period. Before the training of ensemble of SVMs, a stochastic forest of trees for feature selection [Kramer, 2016] was also used. The CNNs used five-second preprocessed windows as input. Feature extraction was coded in MATLAB R2018b. The remaining steps were coded with Spyder 4.0.1 and Python 3.7. All ML functions are from *scikit-learn*. For the CNNs, *TensorFlow* and *Keras* libraries were also used. See the Appendix V for complete details on all pipelines, including the CNNs' architecture.

6.2.3 Developing explanations

After assessing performance, some patients with the following performances were selected: i) high SS and low FPR/h, ii) high SS and high FPR/h, iii) low SS and high FPR/h, and iv) low SS and low FPR/h. It was believed that these represent the whole dataset. By only analysing a reduced number of patients, it was possible to conduct interviews and deepen explanations. It would not be possible to provide

detail on all patients as that would take too many hours per interview. Figure 6.2 depicts an overview of all developed graphical explanations, presented in the interviews.

In the interviews, clinical information was provided, such as seizure onset times, seizure classification, EEG brain activity pattern and vigilance state at seizure onset. On the feature level, the following explanations were presented: i) beeswarm summary plots of SHAP Values [Lundberg and Lee, 2017] (Figure 6.2.b)), which displayed an information-dense summary of how the features impacted the model's output; ii) Partial Dependence Plots (PDPs) [Friedman, 2001] and Individual conditional expectation (ICE) [Goldstein et al., 2015] plots (Figure 6.2.c)), which described the interaction between target response and each input feature; and also iii) logistic regression coefficients (Figure 6.2.d)), for the logistic regression model. The chosen electrodes and features' expected behaviour were also discussed.

A feature-based explanation, inspired by calibration [Vuk and Curk, 2006] and scatter plots, was developed (Figure 6.2.e)). While a calibration curve plots the average predicted probability in each bin (x-axis) against the ratio of positive predictions (y-axis), here, the feature value (x-axis) was plotted against its probability of seizure occurrence (Firing Power value y-axis). In this plot, interictal and preictal samples were coloured differently to inspect features' separability. This way, this plot allowed to visually inspect the classifier behaviour on discriminating individual interictal and preictal windows, and to observe their temporal relation due to the Firing Power values.

The remaining explanations focused on the time series. The Firing Power was plotted along with a sleep/awake detector [Oliveira, 2021], interictal and preictal periods, raised alarms and time-stamps (Figure 6.2.f)). These plots provided insights into classifier dynamics concerning circadian and sleep-wake cycles and changed their configuration concerning each used model. For the logistic regression model, the Firing Power was plotted in black. For the 15 SVMs and three CNNs, all models' Firing Power were plotted in grey and the voting decision in black. Then, for moments considered to be of interest (Figure 6.2.g)), the respective points were plotted over the Firing Power scatter plots and provided counterfactual explanations [Wachter et al., 2017]. Counterfactual explanations are intuitive as these describe a causal situation in the form of "if X had not occurred, Y would not have occurred" [Molnar, 2019]. These explanations were computed by finding the slightest change in each feature that modified the prediction.

For the CNN, Local interpretable model-agnostic explanations (LIME) [Ribeiro et al., 2016] were used to show the points in the EEG time series that made the neural network classify some segments as preictal (Figure 6.2.h)). These explanations were only shown to clinicians as their analysis requires EEG expert knowledge. Although clinicians cannot identify brain patterns that lead to seizures several minutes beforehand, it was assumed they might provide a physiological interpretation

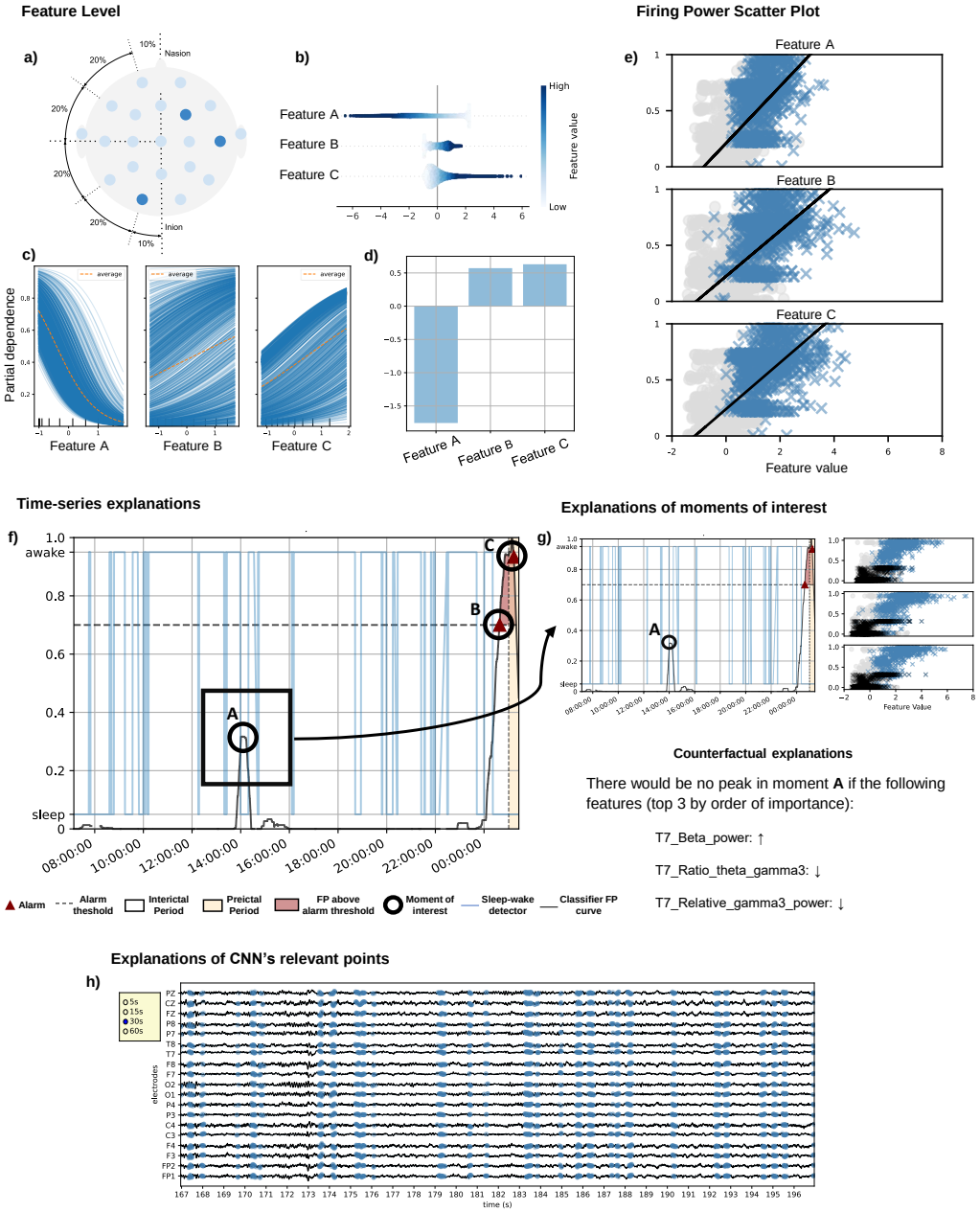


Figure 6.2: The developed explanations for each patient. The selected features were shown and discussed: their expected added value and brain localisation (a). Each feature's SHapley Additive exPlanations (SHAP) value was presented (b), as well as the ICE and PDP (c). Regression coefficients were shown for the logistic regression model (d). A Firing Power scatter plot was also presented (e), where the y-axis represents a seizure probability occurrence (Firing Power) and the x-axis the feature value. The preictal samples are in blue, and the interictal ones are in grey. The Firing Power plot over the signal (f). This plot also shows the sleep/awake patient state and raised alarms. From all signals, some moments that deserved more attention were chosen, mostly false alarms and seizures that the models failed to predict. These moments were analysed (g) by inspecting the correspondent points on the Firing Power scatter plot and developing counterfactual explanations. When using the CNN, in points of interest, Local Interpretable Model-Agnostic Explanations (LIME) was used to inspect the points that led the network to classify the signal as preictal (h).

of the algorithm behaviour, which might correlate with pre-seizure brain dynamics.

All explanations were developed with Spyder 4.0.1 and Python 3.7, along with *sci-kit learn*, *shap*, and *lime* packages.

6.2.4 Evaluating explanations

The developed explanations were shown during interviews while asking open-ended questions. These were shown first to ten data scientists to guarantee that any technical question was addressed. Then, with their feedback, the set of explanations was redefined and presented to ten clinicians. The interviews were audio recorded. The interview script can be found in Appendix X and the presentation slides of the interviews are provided in the Github page. After transcribing and anonymising them into text, the audio files were deleted. All the interviewees agreed to be recorded. The transcribed interviews were analysed using GT, an inductive process broadly used in qualitative methods for social sciences [Chapman et al., 2015]. GT was not used to search traditional themes but to understand how researchers can provide convincing and adequate explanations that allow a clinical translation of these algorithms. Obtained findings are summarised in lessons, as it was believed to best convey the important information.

Note that, despite ten interviewees for each group (data scientists and clinicians) may appear to be a small number, qualitative studies may not require large numbers due to the saturation principle. A strategy based on GT stops when reaching saturation or, in other words, when no more emerging themes and relations can be found [Vasileiou et al., 2018].

6.3 Results

Table 6.1 presents the seizure prediction results for the entire set of patients and patients selected to develop explanations. In Appendix W, the complete results are provided. The obtained five lessons are now explained, each in a different subsection.

6.3.1 Explain your system at different levels

Researchers should divide and order their explanation reasoning into levels according to an increasing granularity or, in other words, from feature and model levels to explaining specific events.

- **Feature level:** show and discuss important features, namely their expected added value, time-window, and correspondent electrode. Also, show each feature's model influence quantitatively. Feature influence was provided through beeswarm summary plots of SHAP values, PDP and ICEs. The regression coefficients were also analysed for the logistic regression model. The models'

Table 6.1: Overall prediction results for the three ML pipelines. The patients that we selected for developing explanations are also present.

Model	Patient	SS (0-1)	FPR/h	Above chance
Logistic Regression	Overall	0.13±0.26	0.40±0.46	5 in 40 (0.125)
	8902	1.00	0.11	Yes
	93402	1.00	0.50	Yes
	101702	0.00	0.71	No
	402	0.00	0.00	No
Ensemble of 15 SVMs	Overall	0.17±0.28	0.87±1.11	7 in 40 (0.175)
	53402	1.00	0.22	Yes
	59102	0.00	0.52	No
	46702	0.00	0.00	No
Ensemble of 3 CNNs	Overall	0.04±0.10	0.18±0.26	3 in 40 (0.075)
	8902	0.50	0.00	Yes
	23902	0.00	1.65	No
	32702	0.00	0.00	No

hyperparameters, such as the support vectors from the trained SVM, could have also been analysed, as suggested during the interviews.

- **Model level:** show how well the model differentiates independent data samples. Graphical explanations are more appealing than only providing fitting metrics and confusion matrixes. An adapted version of calibration and scatter plots was depicted to present the data points' geometric distribution and model discrimination behaviour.

- **Overall system functioning:** provide a visual overview of the system functioning across time and integrate complementary information. Besides providing performance metrics, such as SS and FPR/h, a visual overview provided more information, such as the alarm distribution and the Firing Power over time. Information such as the sleep-wake cycle and time-stamps was also provided, which helped to interpret classifier decisions.

- **System functioning over specific moments:** provide deeper explanations concerning potentially decisive moments. Explanations about all moments could not be inspected as the used recordings were long. People tend only to pay attention and inspect some moments, particularly when models fail: i) false alarms, ii) not predicted seizures, or iii) Firing Power peaks that almost led to false alarms.

6.3.2 Discussing features is important

Discussion about the extracted and selected features occupied a significant part of the interviews' time. Several data scientists asked for more clinical knowledge concerning these features. Also, many wanted to visualise the time plots of some features, particularly spectral bands' relative powers.

Spectral bands' relative power was the most discussed feature. Although these were the most appropriate measures to discuss with clinicians, there were differences in their conceptualisation between data scientists and clinicians. While clinicians obtain spectral band power by visually inspecting the EEG and looking for spikes, ripples, and other abnormal transients, data scientists use mathematical tools, such as Fourier decomposition. Despite these differences in their computation concerning the clinician interpretation, spectral bands' relative power enables extracting relevant information from long-term recordings.

Figure 6.3.b) depicts the importance of discussing features. It shows the selected features and their influence on patient 8902 for the logistic regression model. As one can see, gamma band-related measures appeared in five out of the seven selected features. Clinicians found this suspicious as scalp EEG does not fully capture gamma rhythms. Thus, this predominance of gamma features might be explained by the presence of artefacts. The following hypothesis was discussed: *patient 8902 presents movement jerks caused by EEG pre-seizure dynamics.* Note that this was just a hypothesis as there was no access to video-EEG. The subsection *Making and testing patient-specific conjectures* shows how this hypothesis was tested without accessing video-EEG.

6.3.3 Time plots are the most intuitive explanations

Only showing time plots is not enough, but they are great for raising questions and hypotheses. By inspecting the time plots in Figure 6.3, some patterns were observed, allowing to present hypotheses to clinicians (as explanations).

For example, one may try to explain the false alarm (moment B) after midnight before seizure #4 from patient 8902 by inspecting Figure 6.3.a). When visualising the Firing Power from 8am until midnight, despite presenting a relatively small peak of 0.40 (moment A), it was verified that the system was far from raising an alarm. Afterwards, the Firing Power presented a monotonically increase until reaching a maximum peak value of 1.0 (moment C) during the preictal period. After midnight, the system raised a false alarm (moment B) when it reached a value of 0.7. Despite being a false alarm, this behaviour was considered normal in the light of the preictal period assumption: a transitional stage between seizure-free brain activity and a seizure, able to be captured from an EEG background analysis [Pinto et al., 2022, Scheffer et al., 2009, Litt et al., 2001]. All interviewees accepted that the system raised a false alarm because this brain transition may have started minutes earlier than the interval considered when training the model.

By visualising the data from seizure #4 of patient 93402 (Figure 6.3.c)), three distinct patterns of alarms over time were identified: between midnight and 6am (red circles), around mid-day (yellow circles), and around 6pm (green circles). Additionally, in this patient, both testing seizures occurred between midnight and 6am

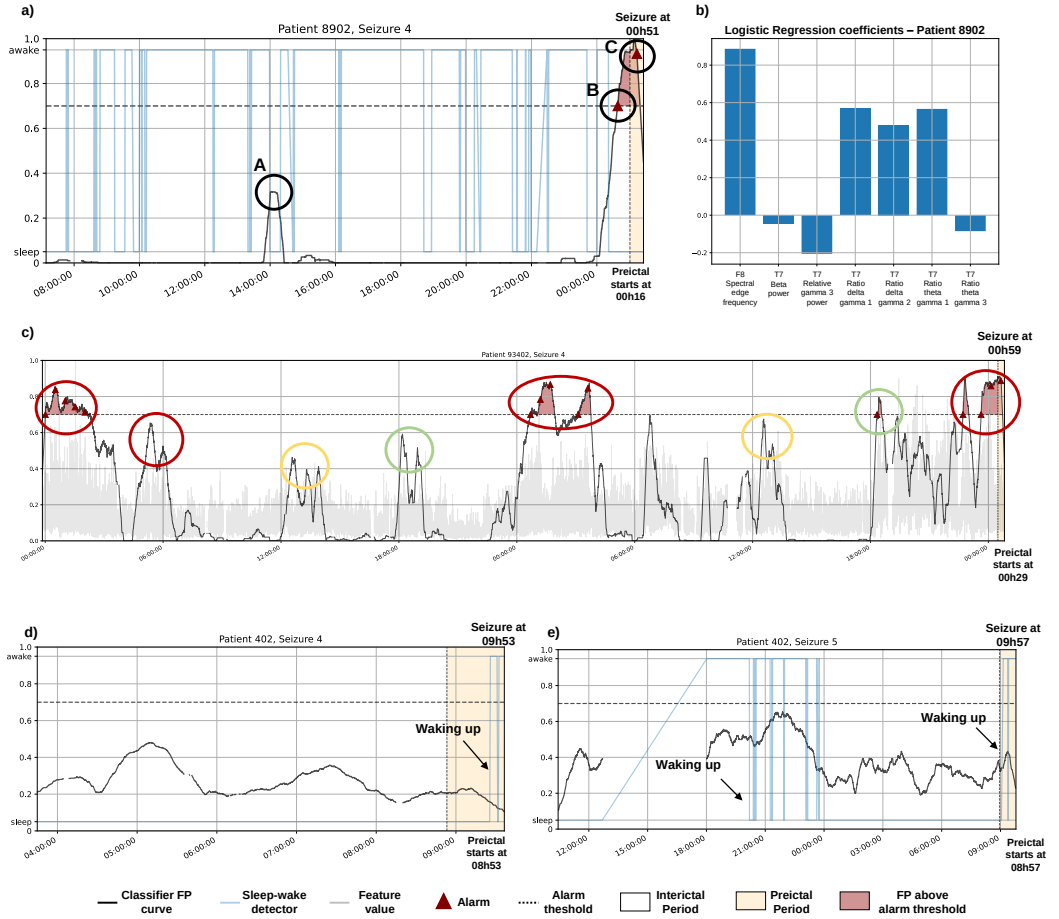


Figure 6.3: Examples of explanations for the logistic regression model. The time-series explanation (a) and the corresponding feature regression coefficients (b) are presented for the seizure #4 of patient 8902. The Firing Power and one feature (wavelet energy of the decomposition level 5 from electrode O5) are plotted over time for seizure #4 of patient 93402 (c). By inspecting each feature individually, very straightforward relations were observed between the models' decisions and the corresponding feature. Each coloured circles represent similar signal activity captured at the same time of day (different colors corresponding to repeating patterns showing at different times of the 24h day), possibly evidencing the influence of circadian rhythms. Time series explanations are also presented when no alarm was raised, for seizures #4 (d) and #5 (e) of patient 402.

(seizure #4 at 01:35 and seizure #5 at 06:10), which fit into the red cluster. The following hypothesis was presented: *these false alarm clusters suggest the existence of periods of brain susceptibility to seizures, which may not always lead to seizures.* This rationale is a paradigm shift from prediction to forecasting [Dumanis et al., 2017], which intuitively appeared during interviews. A Firing Power increase over days was observed in the yellow and green circles, which suggests an effect of medication tapering, as told by one clinician. These hypotheses might explain the occurrence of false positives.

Other hypotheses arose when observing the time plots. In patient 8902, could the selected gamma-related features (Figure 6.3.b)) be related to sleep-wake states? There are significant differences as the wake EEG contains more muscle artefacts

than the sleep one, which might introduce a strong bias.

When inspecting Figures 6.3.d) and 6.3.e) (seizures #4 and #5 from patient 402), one may hypothesise that around the waking up time, there is a higher risk of seizure (both testing seizures occur at waking up). For these cases, the logistic regression model never raised any alarm (see Table 6.1).

All these explanations are merely hypotheses and must be tested. Testing the developed hypothesis is one of the major topics of this paper, which concerns the last lesson.

6.3.4 Interictal and preictal concepts differ between data scientists and clinicians

For a clinician, the preictal period is often considered a fast spontaneous phenomenon that might last for less than one second and start some seconds before the seizure onset. Data scientists' strategy of classifying consecutive windows followed by regularisation [Teixeira et al., 2012] while trying to capture a transitional background state into a seizure one does not hold the clinicians' notion of seizure generation.

Also, the interictal period is considered by data scientists as the period between postictal and preictal periods of subsequent seizures. To clinicians, it concerns the periods that are not postictal, preictal, or ictal, and that necessarily contains abnormal epileptiform activity. In other words, not all data considered by data scientists as interictal matches the one which is considered by clinicians.

Understanding these differences is extremely important, as ML model development is based on the provided data. Data scientists deliver data samples to ML algorithms, in which a sample is usually a 5-second window or features computed in that window. Each window is labelled as either preictal or interictal. Then, the ML algorithm trains a model that best discriminates these data samples. Furthermore, other technical aspects also need to be considered. Although the data scientists' preictal period may be more extended than the clinicians' one (a fixed period of usually several minutes until some hours before a seizure), it is still a rare event causing an extreme data imbalance. Also, although this approach is established in literature, it is hard for clinicians to understand that data scientists must define a fixed preictal period for each patient.

Figure 6.4 shows the ensemble of SVMs for patients 53402, 46702, and 59102. These plots show a strong agreement between the ensemble voting (black line) and each SVM's decisions (grey lines). Each SVM model was different as the feature selection and data balance steps were stochastic operations. Clinicians and data scientists were asked if this similarity was a good sign. The data scientists were quite satisfied since the SVM output was reasonably coherent between each execution. Clinicians were not consensual, and they tended to refrain from answering due to the data scientists' concept of interictal period.

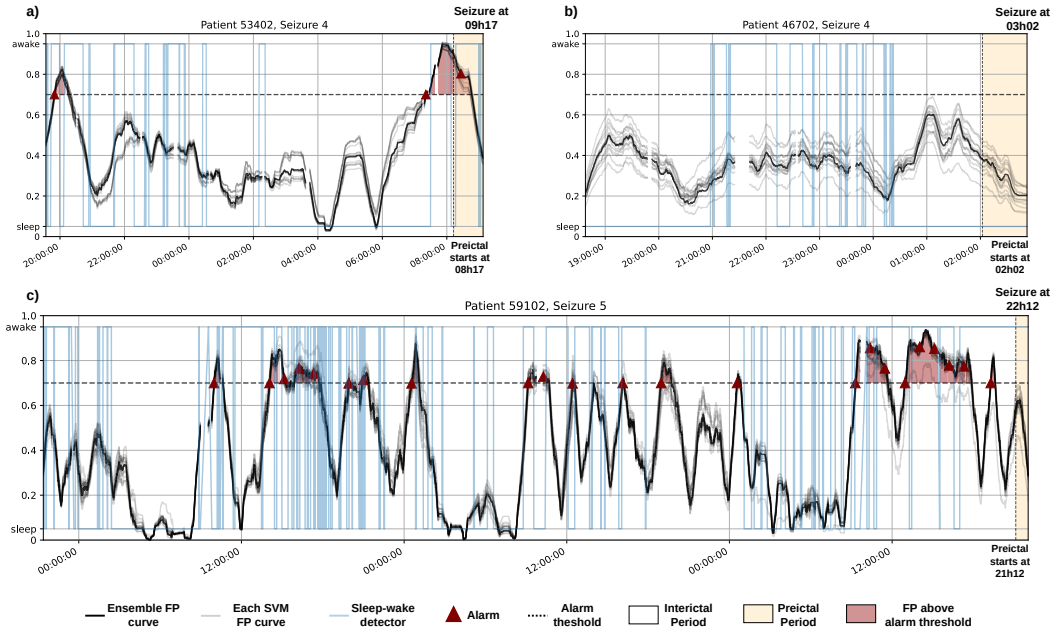


Figure 6.4: Plot of the ensemble of SVMs' decisions over time for different patients: patient 53402 seizure #4 (a), patient 46702 seizure #4 (b), and patient 59102 seizure #5 (c). Each SVM decision is in grey, and the ensemble voting system is in black.

6.3.5 Making and testing conjectures is the solution to explain a ML model decision when there is no solid physiology grounding

If clinicians cannot explain pre-seizure mechanisms several minutes (or even hours [Litt et al., 2001]) before seizure onset, data scientists will probably not be the ones doing it as they cannot provide a clinician-comprehensible answer. One clinician stated a critical idea: "until proven otherwise, everything is an artefact", which may be essential to understanding the utility of explainability. For seizure prediction, the goal of explainability is not merely to explain decisions but to develop hypotheses based on physiology mechanisms, which must be tested. When researchers reject the null hypotheses, they gain insight and trust. If the models fail, study assumptions must be reviewed and the used methodologies and explanations redesigned, leading to a more complete problem formalisation. The loop continues until one can trust the obtained methods. This rationale is depicted in Figure 6.5.

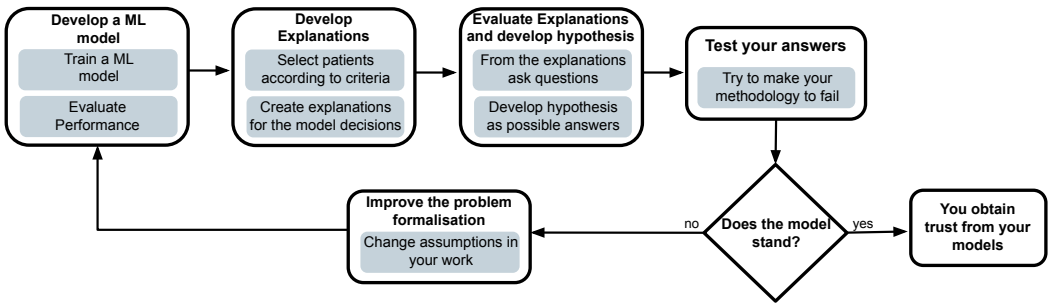


Figure 6.5: The workflow process of explaining ML models.

6.3.5.1 Making and testing patient-specific conjectures

As mentioned, patient 8902's logistic regression model achieved high performance using several gamma-related features, which provoked scepticism as scalp EEG was used and, therefore, it would be difficult to capture gamma activity. Two hypotheses arise about the gamma features: i) there might be predictive power in the gamma features related to the GABAergic system and a high-frequency synchrony increase, or ii) gamma-related features might have captured muscle artefacts as patient 8902 presented muscle jerks which resulted from pre-seizure dynamics. Concerning hypothesis i), it cannot be stated and tested because there are no intracranial EEG recordings from this patient, to compare it.

Concerning hypothesis ii), there is the possibility of visually inspecting the EEG windows where gamma spectral-band power is significant to gain some insight, but that would be arduous due to the necessity of experts to analyse hours of data. It would also be possible to search for tools that classify signal segments into artefacts, noise, or EEG-related phenomena, which would also take time. Here, a faster strategy was followed. To understand the gamma-band contribution, the logistic regression models were retrained six times where, for each time, one of the following features was used: spectral-band power features from either delta, theta, beta, or alpha bands, signal's time variance, and signal's total band power. Time variance was used to understand if time-frequency features could also capture similar dynamics.

For seizure #4 (Figure 6.6.a)), all spectral bands, total spectral power, and time signal variance had a similar discriminatory capacity. With seizure #5 (6.6.b)), despite the majority of the features' models presenting a similar morphology, there were differences. Delta, theta, and total power models would raise more false alarms, while gamma and alpha bands would not raise any. The beta band power and time variance presented differences in their Firing Power dynamics, where the first often exceeded the alarm threshold. These findings mitigated some of the gamma-band features' scepticism, and the muscle-jerk hypothesis was rejected. As the same patterns appeared in most spectral bands, total spectra, and within the temporal domain, they suggest a general EEG background change [Scheffer et al., 2009, Litt et al., 2001] due to pre-seizure dynamics.

For patients 93902 and 402, hypotheses related to circadian and sleep-wake cycles could be constructed, respectively. Nevertheless, there is the need to use more data to test these hypotheses. As these hypotheses could not be tested, they cannot be truly stated. However, one may include other strategies to speed up this process: the clinicians pointed out their curiosity to understand how the models would behave when patients performed daily activities, such as eating, getting up, and scratching their heads. This rationale relates to trying to make the methodology fail. If there was a way to guarantee that the models would not confound these activities with a pre-seizure state, trust would increase. Also, when developing methods to detect

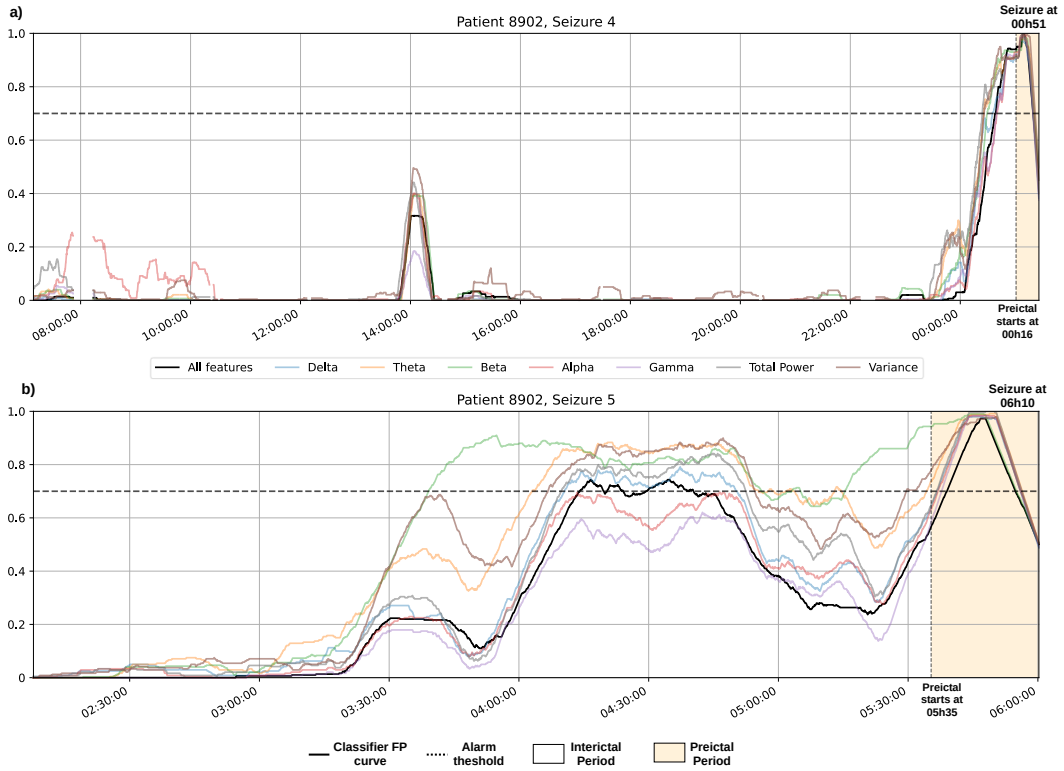


Figure 6.6: Patient 8902's Firing Power time plots for different logistic regression models for both testing seizures (a and b). The black line represents the original model, the remaining concern logistic regression models trained with only a determined spectral band power, total band power, or the signal time variance.

such activities, one may better understand the circadian cycle's influence.

6.3.5.2 Making and testing patient-general conjectures

Hypotheses were also developed and tested from the entire set of patients. By inspecting all patients' Firing Power plots using the logistic regression model, several patients presented, for at least one seizure, one of the following scenarios: i) the model could not predict a seizure, but one could trust its behaviour when inspecting the time plot; or ii) the model's Firing Power behaviour was poor, but one could find a sleep-wake transition in the preictal period, suggesting a sleep-wake cycle influence (as patient 402 in Figure 6.3). There were 14 and 21 patients for the first and second scenarios. Additionally to patient 93402, there were four more patients (59102, 95202, 109502, 112802) (Figure 6.3) where false alarms cycles occurred within similar periods in consecutive days, suggesting a circadian-cycle influence. See Appendix Y for examples to better understand these cases and the Github page for all patients' time plots.

Lastly, clinicians asked about a possible performance bias towards patients whose training and testing seizures occurred around similar times of the day. Thus, a fore-

casting rationale was applied to the logistic regression models, where high seizure-risk warnings corresponded to Firing Power values over the alarm threshold. These were then compared to a circadian forecasting algorithm that only used circadian information. This procedure was only performed for the logistic regression models for simplicity reasons. For each tested seizure, the circadian algorithm raised high seizure-risk warnings from 30 minutes before to 30 minutes after each seizure training onset time (see Appendix Z for an illustration of this algorithm). By comparing seizure sensitivities and Time in Warning (TiW), it was verified that the EEG-based models outperformed the circadian forecasting ones in both metrics. The EEG-based models obtained a sensitivity of 0.29 and a TiW of 1h32, while the circadian forecasting obtained a sensitivity of 0.15 and a TiW of 2h52. Moreover, six EEG-based patient models presented simultaneously higher sensitivity and lower TiW values. See Appendix Z for the full comparison results between the two approaches.

All the counts in this subsection have statistical significance based on accumulative binomial distribution tests (α=0.05) [Pinto et al., 2022, Pinto et al., 2021a, Alvarado-Rojas et al., 2014]. The complete statistical analysis is detailed in Appendix Y.

6.4 Discussion

Although state-of-the-art features were used to characterise the EEG signal, identifiy manifestations of the pre-seizure state, and inspect the models' Firing Power, these explanations were insufficient to understand brain dynamics of seizure generation. ML explainability might not be the proper tool to understand the EEG, which justifies research efforts on developing specific strategies to make their models fail. It is difficult for clinicians to reject the developed models when they hold against a systematic conjecture testing. Thus, authors should be conservative in their explanations and base model development on an iterative refutation algorithm when the underlying physiologic mechanisms are unknown.

This reasoning relates to Karl Popper's falsification theory, in which conjecture and attempted falsification should drive model development [Taran et al., 2021,Forde and Paganini, 2019, Shahar, 1997]. It is essential to review these concepts as researchers tend to intrinsically demonstrate that seizure prediction is possible solely based on testing models. In other words, current studies test the research hypothesis directly, relying on inductive reasoning instead of trying to reject a null hypothesis. Although inductive logic is valuable and practical, it is prone to confirmation bias, easily achieved in ML problems due to overfitting and overtesting. Curiously, the problem of induction, initially stated in David Hume's Uniformity of Nature Assumption [Wilkinson, 2013], stands until today.

The lack of understanding on the ictogenesis process was addressed by showing the EEG to clinicians. During the interviews, the EEG segments marked as relevant

were shown to them along with the data samples that the CNNs found relevant in predicting seizures. This task was challenging as clinicians could not find any pattern. Very often, the CNN model selected some points as relevant and, some seconds after, did not select similar points. Thus, the CNN classifier did not use criteria that followed a neurologist's rationale. A long-term EEG monitorisation may not lead to a higher understanding of the ictogenesis process. However, it can be beneficial to capture factors that may influence a higher susceptibility to seizures (seizure forecasting [Stirling et al., 2021a, Karoly et al., 2020, Karoly et al., 2017]).

For the case of seizure prediction, this work proves that a transparent model, such as the case of a logistic regression, may not lead to more convincing explanations than the ones using more complex models. No clinician raised any issue concerning model transparency when presenting the SVM and CNN classifiers, which might be due to the mentioned lack of EEG ictogenesis-related knowledge. Despite the literature is heading toward more complex models [Pinto et al., 2021b, Kuhlmann et al., 2018b, Freestone et al., 2017], there is the need to stress this finding as some authors advocate against the use of such models due to their black-box nature [Rudin, 2019] or due to the General Data Protection Regulation (GDPR) Right to an Explanation (article 22) [Goodman and Flaxman, 2017, Kaminski, 2019], which is still purposely vague.

Explaining seizure prediction models led to considering each case scenario individually. Different hypotheses were considered for each patient, where their conjecture testing was made differently. When verifying all patients' time plots, typical model behaviours were found, which may be a consequence of epilepsy's clinical heterogeneity. Also, for a significant number of patients, the EEG-based models outperformed (in both terms of SS and TiW) the circadian forecasting models. Although the used data does not concern real-life, this finding confirms the added value of developing EEG-based prediction approaches for some patients.

6.4.1 Translation to other healthcare problems

This work may translate to other ML healthcare applications, particularly EEG-related, where physiological mechanisms that generate an event to be predicted are not fully understood. For instance, predicting hospital mortality after acute coronary events [Valente et al., 2021, Granger et al., 2003] might require a higher degree of transparency as there are established score models grounded on a clinical rationale. Some lessons might need further adaptation to each problem, where these can be stated in a generalised form. For example, there is possible to translate the conceptual difference between interictal and preictal periods to: *carefully review the used rationale to label data into categories and discuss it with clinicians.*

6.4.2 Study limitations

This work had several limitations. It is hard to provide extensive examples of some conjecture testing strategies, as the used data comprised presurgical monitoring conditions. Some explanation hypotheses require new testing data and extensive recording periods to capture many specific events. Although hypotheses have been made about circadian and sleep-wake cycles' influence, which the literature supports [Stirling et al., 2021a, Karoly et al., 2020, Karoly et al., 2017], there is not sufficient data to confirm it. It would also be valuable to have access to video-EEG and to find more patients presenting a clear EEG-background transition from a baseline to pre-seizure activity, as patient 8902.

Counterfactual explanations did not reveal to be fundamental but are essential in many applications, as they are gaining prominence within technical, legal, and business circles for ML [Verma et al., 2020, Barocas et al., 2020, Wachter et al., 2017]. This study might have failed to apply counterfactual explanations as these were used to explain changes in EEG dynamics that influenced the models' decisions. These explanations were not pertinent as patients cannot change their brain dynamics. Counterfactual explanations tend to be more useful when the user can intervene in the decision, such as a bank decision on a loan: *if you had done X, you would have got the loan*. Nevertheless, they might be helpful in forecasting when analysing a large quantity of data and accessing information the patient can control, such as medication, sleeping, eating, and other daily activities.

Our lessons result from GT's iterative and emergent observation and analysis. As GT is not suited for verification/falsification of preexisting propositions [Timonen et al., 2018, Cho and Lee, 2014], authors are invited to perform similar methodologies (including other ML healthcare problems and data types). Due to a mostly data science and epilepsy background, there is the need to recognise an increased difficulty in using qualitative research tools [Doshi-Velez and Kim, 2017].

6.4.3 Final reflections

Clinicians do not fully understand EEG pre-seizure mechanisms occurring several minutes before a seizure; nevertheless, it was possible to provide more or less convincing explanations for certain model decisions without requiring fully transparent models.

For predicting seizures and other healthcare ML-related events where clinical apriori knowledge is limited, explainability is not about simply explaining decisions but improving the developed models, reviewing used assumptions and, thus, gaining trust. Basing model development on an iterative refutation algorithm might promote trust while dealing with a lack of clinical grounding.

For future work, there is the need to repeat this methodology in ultra-long term recordings from real-life, such as the study from [Cook et al., 2013]. Other case sce-

narios may be found with ultra-long term data and confirm (or reject) the reported ones. Only a long-term analysis will tell if these explanation methods remain effective when inspecting days and months of data. There is also the need to consider interviewing patients to understand their perspectives, relation to devices, and how to help them deal with the devices' predictions in cases of failure.

Chapter 7

Conclusions

In this thesis, efforts were made to improve the communication between clinicians and researchers and highlight possible paths to clinical translation. This challenge was faced by using a multidisciplinary approach, which joined engineering domain techniques with tools from the social sciences. From the three main contributions, it was possible to retrieve some conclusions.

Based on the analysis of the seizure prediction ecosystem study, four guidelines were proposed for a higher chance of clinical acceptance, where authors should: i) state their assumptions regarding brain dynamics before presenting the mathematical tools used in data analysis; ii) state the prospective applications envisioned with the designed experiment; iii) use methodologies that have been clinically approved as a gold standard for comparison; and iv) focus on explainability to promote trust among clinicians and other data scientists. The guideline that concerns explainability was considered the most important.

The development of an evolutionary framework, whose output is an intrinsically interpretable model, leads to a way to extract patient-specific and patient-general knowledge. Although the models' performance was not acceptable for clinical translation, this methodology may inspire other approaches to extract knowledge and predict rare events from time series. If researchers wish to use this methodology, they may only need to adapt the genotype and the fitness function.

A formal evaluation of explainability towards clinical and data science specialists, which ranged from intrinsically interpretable models to black-box ones, allowed to verify that, when *a priori* knowledge is limited, the goal of explainability is not to make decisions merely but to improve the developed models, review used assumptions and, thus, gain trust. Basing model development on an iterative refutation algorithm might give trust while dealing with a lack of clinical grounding.

Seizure prediction heads towards acquiring ultra-long-term Electroencephalogram (EEG), enabled by recent technology, such as the UNEEG SubQ [Duun-Henriksen et al., 2020, Weisdorf et al., 2019], the ByteFlies Sensor Dots [Nasseri et al., 2020], Epiminder [Stirling et al., 2021c], Percept PC [Gregg et al., 2021],

and RNS system [Sun and Morrell, 2014]. The rationale of this thesis passed by anticipating future problems of researchers having access to this type of data since, for this work, there was no opportunity to access such data. Naturally, there is the need to adapt and test all developed methodologies in ultra-long-term recordings from real life. Other case scenarios may be found with this data which may confirm (or reject) the reported ones. Authors should explore and integrate information concerning concept drifts and cyclical variables, such as medication intake, circadian cycles, sleep, multidien cycles, and daily activities [Baud et al., 2022, Karoly et al., 2020, Kuhlmann et al., 2018b, Baud et al., 2018, Dumanis et al., 2017].

The developed methodologies should have also involved interviews with lawyers and patients to understand better this research area and the future implications of prediction and forecasting devices. Understanding the patients' perspectives towards these devices, particularly their reaction to when they fail and how to maintain trust, may figure a tremendous challenge [Baud et al., 2022, Bruno et al., 2018, Schulze-Bonhage et al., 2010].

Researchers are transitioning from prediction horizons of minutes (prediction) to probabilistic horizons of hours or days (forecasting) [Proix et al., 2021, Karoly et al., 2020]. Nevertheless, in terms of application to the patient, one might still need to choose a threshold to perform an intervention. Interventions will change concerning the horizon. With forecasting, patients might plan their lives around or undergo a short cycle of benzodiazepine intake instead of a single intake when using prediction (or combining both forecasting and prediction). There is the need to conduct studies on both approaches to compare them to understand the benefit-risk ratio, patient resilience against false interventions and, most importantly, their reaction to when these systems fail [Baud et al., 2022].

Despite inherent difficulties in conducting science and the need to have the means to perform research, authors need to understand their significant bias towards people living in rich and western countries [Taylor and Rommelfanger, 2022, Maina et al., 2021]. In low-income countries, a routine-EEG or other exams (such as Magnetic Resonance Imaging (MRI) a Computed Tomography (CT) scan) may be a luxury that the majority of people cannot afford or may not even be available [Maina et al., 2021, McLane et al., 2015]. Despite these difficulties, devices for epilepsy management will need to become accessible in such countries. The *long and winding road*, as coined by [Mormann et al., 2007], continues.

References

[Aarabi et al., 2009] Aarabi, A., Fazel-Rezai, R., and Aghakhani, Y. (2009). Eeg seizure prediction: measures and challenges. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1864–1867. IEEE. 161

[Acharya et al., 2013] Acharya, U. R., Sree, S. V., Swapna, G., Martis, R. J., and Suri, J. S. (2013). Automated eeg analysis of epilepsy: a review. *Knowledge-Based Systems*, 45:147–165. 14, 57, 59, 164

[Adeli et al., 2007] Adeli, H., Ghosh-Dastidar, S., and Dadmehr, N. (2007). A wavelet-chaos methodology for analysis of eegs and eeg subbands to detect seizure and epilepsy. *IEEE Transactions on Biomedical Engineering*, 54(2):205–211. 14, 53

[Alkan et al., 2005] Alkan, A., Koklukaya, E., and Subasi, A. (2005). Automatic seizure detection in eeg using logistic regression and artificial neural network. *Journal of neuroscience methods*, 148(2):167–176. 40

[Alotaiby et al., 2014] Alotaiby, T. N., Alshebeili, S. A., Alshawi, T., Ahmad, I., El-Samie, A., and Fathi, E. (2014). Eeg seizure detection and prediction algorithms: a survey. *EURASIP Journal on Advances in Signal Processing*, 2014(1):1–21. 14, 174

[Alvarado-Rojas et al., 2014] Alvarado-Rojas, C., Valderrama, M., Fouad-Ahmed, A., Feldwisch-Drentrup, H., Ihle, M., Teixeira, C., Sales, F., Schulze-Bonhage, A., Adam, C., Dourado, A., et al. (2014). Slow modulations of high-frequency activity (40–140 hz) discriminate preictal changes in human focal epilepsy. *Scientific reports*, 4(1):1–9. 26, 50, 52, 54, 56, 58, 59, 61, 93, 100, 121, 256

[Anastasiadou et al., 2019] Anastasiadou, M. N., Christodoulakis, M., Papathanasiou, E. S., Papacostas, S. S., Hadjipapas, A., and Mitsis, G. D. (2019). Graph theoretical characteristics of eeg-based functional brain networks in patients with epilepsy: The effect of reference choice and volume conduction. *Frontiers in Neuroscience*, 13:221. 16

[Andrzejak et al., 2009] Andrzejak, R. G., Chicharro, D., Elger, C. E., and Mormann, F. (2009). Seizure prediction: any better than chance? *Clinical Neurophysiology*, 120(8):1465–1478. 31, 34

[Andrzejak et al., 2003] Andrzejak, R. G., Mormann, F., Kreuz, T., Rieke, C., Kraskov, A., Elger, C. E., and Lehnertz, K. (2003). Testing the null hypothesis of the nonexistence of a preseizure state. *Physical Review E*, 67(1):010901. 34, 92, 101, 109, 165, 180

[Anyanwu and Motamedi, 2018] Anyanwu, C. and Motamedi, G. K. (2018). Diagnosis and surgical treatment of drug-resistant epilepsy. *Brain sciences*, 8(4):49. 13

[Apley and Zhu, 2020] Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086. 42, 165, 182

[Assi et al., 2017] Assi, E. B., Nguyen, D. K., Rihana, S., and Sawan, M. (2017). Towards accurate prediction of epileptic seizures: A review. *Biomedical Signal Processing and Control*, 34:144–157. xxii, 25, 30, 34, 36, 48, 49, 53, 54, 57, 58, 59, 76, 78, 83, 87, 109, 161, 163, 164, 166, 174, 176, 178, 179, 181

[Assi et al., 2015] Assi, E. B., Sawan, M., Nguyen, D. K., and Rihana, S. (2015). A hybrid mrmr-genetic based selection method for the prediction of epileptic seizures. In *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4. IEEE. 50, 51, 52, 56, 57, 58, 60, 162

[Bäck et al., 2020] Bäck, T., Preuss, M., Deutz, A., Wang, H., Doerr, C., Emmerich, M., and Trautmann, H. (2020). *Parallel Problem Solving from Nature-PPSN XVI: 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part II*, volume 12270. Springer Nature. 165, 182

[Baghdadi et al., 2021] Baghdadi, A., Daoud, S., Dammak, M., Mhiri, C., Siarry, P., Alimi, A. M., et al. (2021). A channel-wise attention-based representation learning method for epileptic seizure detection and type classification. *IEEE Journal of Biomedical and Health Informatics*. 63, 64

[Bai et al., 2015] Bai, Y., Liang, Z., and Li, X. (2015). A permutation lempel-ziv complexity measure for eeg analysis. *Biomedical Signal Processing and Control*, 19:102–114. 87

[Bandarabadi et al., 2015a] Bandarabadi, M., Rasekhi, J., Teixeira, C. A., Karami, M. R., and Dourado, A. (2015a). On the Proper Selection of Preictal Period for Seizure Prediction. *Epilepsy and Behavior*, 46:158–166. 3, 57, 164

[Bandarabadi et al., 2012] Bandarabadi, M., Teixeira, C. A., Direito, B., and Dourado, A. (2012). Epileptic seizure prediction based on a bivariate spectral power methodology. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5943–5946. IEEE. 57, 59

[Bandarabadi et al., 2015b] Bandarabadi, M., Teixeira, C. A., Rasekhi, J., and Dourado, A. (2015b). Epileptic seizure prediction using relative spectral power features. *Clinical Neurophysiology*, 126(2):237–248. 14, 36, 50, 51, 52, 53, 54, 56, 58, 87, 100, 104, 162, 164

[Barocas et al., 2020] Barocas, S., Selbst, A. D., and Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89. 123

[Bartz-Beielstein et al., 2014] Bartz-Beielstein, T., Branke, J., Mehnen, J., and Mersmann, O. (2014). Evolutionary Algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(3):178–195. 84

[Baud et al., 2018] Baud, M. O., Kleen, J. K., Mirro, E. A., Andrechak, J. C., King-Stephens, D., Chang, E. F., and Rao, V. R. (2018). Multi-day rhythms modulate seizure risk in epilepsy. *Nature communications*, 9(1):1–10. 2, 27, 62, 126, 165, 174

[Baud et al., 2022] Baud, M. O., Proix, T., Gregg, N. M., Brinkmann, B. H., Nurse, E. S., Cook, M. J., and Karoly, P. J. (2022). Seizure forecasting: bifurcations in the long and winding road. *Epilepsia*. 24, 126

[Baud et al., 2020] Baud, M. O., Proix, T., Rao, V. R., and Schindler, K. (2020). Chance and risk in epilepsy. *Current opinion in neurology*, 33(2):163–172. 23, 27, 37

[Bautista and Glen, 2009] Bautista, R. E. D. and Glen, E. T. (2009). Seizure severity is associated with quality of life independent of seizure frequency. *Epilepsy & Behavior*, 16(2):325–329. 9

[Becker et al., 2020] Becker, T., Vandecasteele, K., Chatzichristos, C., Van Paesschen, W., Valkenborg, D., Van Huffel, S., and De Vos, M. (2020). Classification with a deferral option and low-trust filtering for automated seizure detection. *Sensors*. 166, 176, 193

[Beckers et al., 2021] Beckers, R., Kwade, Z., and Zanca, F. (2021). The eu medical device regulation: Implications for artificial intelligence-based medical device software in medical physics. *Physica Medica*, 83:1–8. 2, 166, 185

[Becris, 2021a] Becris (2021a). Becris Icon made by Flat Icon from www.flaticon.com. FlaIcon Website in https://https://www.flaticon.com. xxiii, 69

[Becris, 2021b] Becris (2021b). Neural Icon made by Becris from www.flaticon.com. FlaIcon Website in https://https://www.flaticon.com. xxiii, 69

[Ben-Menachem, 2002] Ben-Menachem, E. (2002). Vagus-nerve stimulation for the treatment of epilepsy. *The Lancet Neurology*, 1(8):477–482. 166, 183

[Beniczky et al., 2021] Beniczky, S., Karoly, P., Nurse, E., Ryvlin, P., and Cook, M. (2021). Machine learning and wearable devices of the future. *Epilepsia*, 62:S116–S124. 23, 24

[Berry et al., 2012] Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Marcus, C., Vaughn, B. V., et al. (2012). The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176:2012. 38

[Bialer et al., 2017] Bialer, M., Johannessen, S. I., Levy, R. H., Perucca, E., Tomson, T., White, H. S., and Koepp, M. J. (2017). Seizure detection and neuromodulation: A summary of data presented at the xiii conference on new antiepileptic drug and devices (eilat xiii). *Epilepsy research*, 130:27–36. 27

[Bigelow and Kouzani, 2019] Bigelow, M. D. and Kouzani, A. Z. (2019). Neural stimulation systems for the control of refractory epilepsy: a review. *Journal of NeuroEngineering and Rehabilitation*, 16(1):1–17. xxix, 21, 22

[Biondi et al., 2022] Biondi, A., Santoro, V., Viana, P. F., Laiou, P., Pal, D. K., Bruno, E., and Richardson, M. P. (2022). Noninvasive mobile eeg as a tool for seizure monitoring and management: A systematic review. *Epilepsia*, 63(5):1041–1063. 17

[Blume et al., 2015] Blume, C., Del Giudice, R., Wislowska, M., Lechinger, J., and Schabus, M. (2015). Across the consciousness continuum—from unresponsive wakefulness to sleep. *Frontiers in human neuroscience*, 9:105. xxii, 38

[Boddu and Kumari, 2020] Boddu, S. H. and Kumari, S. (2020). A short review on the intranasal delivery of diazepam for treating acute repetitive seizures. *Pharmaceutics*, 12(12):1167. xxix, 24

[Boon et al., 2018] Boon, P., De Cock, E., Mertens, A., and Trinka, E. (2018). Neurostimulation for drug-resistant epilepsy: a systematic review of clinical evidence for efficacy, safety, contraindications and predictors for response. *Current opinion in neurology*, 31(2):198–210. 21, 23

[Boon et al., 2007] Boon, P., Vonck, K., De Herdt, V., Van Dycke, A., Goethals, M., Goossens, L., Van Zandijcke, M., De Smedt, T., Dewaele, I., Achten, R., et al. (2007). Deep brain stimulation in patients with refractory temporal lobe epilepsy. *Epilepsia*, 48(8):1551–1560. 166, 183

[Borgelt and Kruse, 2002] Borgelt, C. and Kruse, R. (2002). Induction of association rules: Apriori implementation. In *Compstat*, pages 395–400. Springer. 84, 93, 97

[Bouw et al., 2021] Bouw, M. R., Chung, S. S., Gidal, B., King, A., Tomasovic, J., Wheless, J. W., and Van Ess, P. J. (2021). Clinical pharmacokinetic and pharmacodynamic profile of midazolam nasal spray. *Epilepsy Research*, 171:106567. xxix, 24

[Boyatzis, 1998] Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. sage. 67

[Briden and Norouzi, 2021] Briden, M. and Norouzi, N. (2021). 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)*, pages 1092–1095. IEEE. 63

[Brinkmann et al., 2022] Brinkmann, B., Nurse, E., Nasseri, M., Viana, P. F., Karoly, P., Attia, T. P., Gregg, N., Joseph, B., Grzeskowiak, C., Dümpelmann, M., et al. (2022). Seizure forecasting and detection with wearable devices and subcutaneous eeg-a practical seizure gauge (n2. 004). 24

[Brinkmann et al., 2021] Brinkmann, B. H., Karoly, P. J., Nurse, E. S., Dumanis, S. B., Nasseri, M., Viana, P. F., Schulze-Bonhage, A., Freestone, D. R., Worrell, G., Richardson, M. P., et al. (2021). Seizure diaries and forecasting with wearables: epilepsy monitoring outside the clinic. *Frontiers in Neurology*, 12. 62

[Bruno et al., 2018] Bruno, E., Simblett, S., Lang, A., Biondi, A., Odoi, C., Schulze-Bonhage, A., Wykes, T., Richardson, M. P., Consortium, R.-C., et al. (2018). Wearable technology in epilepsy: the views of patients, caregivers, and healthcare professionals. *Epilepsy & Behavior*, 85:141–149. 126

[Bulusu et al., 2021] Bulusu, S., Prasad, R. S. S. S., Telluri, P., and Neelima, N. (2021). Methods for epileptic seizure prediction using eeg signals: A survey. In *Artificial Intelligence Techniques for Advanced Computing Applications*, pages 101–115. Springer. 87

[Chapman et al., 2015] Chapman, A., Hadfield, M., and Chapman, C. (2015). Qualitative research in healthcare: an introduction to grounded theory using thematic analysis. *Journal of the Royal College of Physicians of Edinburgh*, 45(3):201–205. 67, 70, 113

[Charmaz and Belgrave, 2007] Charmaz, K. and Belgrave, L. L. (2007). Grounded theory. *The Blackwell encyclopedia of sociology*. 70

[Chen and Cherkassky, 2020] Chen, H.-H. and Cherkassky, V. (2020). Performance metrics for online seizure prediction. *Neural Networks*, 128:22–32. 26

[Chiang et al., 2022] Chiang, S., Baud, M. O., Worrell, G. A., and Rao, V. R. (2022). Seizure forecasting and detection: Computational models, machine learning, and translation into devices. *Frontiers in Neurology*, page 444. 39

[Chisci et al., 2010] Chisci, L., Mavino, A., Perferi, G., Sciandrone, M., Anile, C., Colicchio, G., and Fuggetta, F. (2010). Real-time epileptic seizure prediction using ar models and support vector machines. *IEEE Transactions on Biomedical Engineering*, 57(5):1124–1132. 34, 51, 53, 162

[Cho and Lee, 2014] Cho, J. Y. and Lee, E.-H. (2014). Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *Qualitative report*, 19(32). 123

[Chokroverty, 2009] Chokroverty, S. (2009). *Sleep Disorders Medicine E-Book: Basic Science, Technical Considerations, and Clinical Aspects*. Elsevier Health Sciences. 38

[Chokroverty, 2017] Chokroverty, S. (2017). Overview of normal sleep. In *Sleep disorders medicine*, pages 5–27. Springer. 38

[Chu et al., 2016] Chu, L. F., Utengen, A., Kadry, B., Kucharski, S. E., Campos, H., Crockett, J., Dawson, N., and Clauson, K. A. (2016). "nothing about us without us"—patient partnership in medical conferences. *BMJ*, 354. 81, 166

[Cloppenborg et al., 2016] Cloppenborg, T., May, T. W., Blümcke, I., Grewe, P., Hopf, L. J., Kalbhenn, T., Pfäfflin, M., Polster, T., Schulz, R., Woermann, F. G., et al. (2016). Trends in Epilepsy Surgery: Stable Surgical Numbers Despite Increasing Presurgical Volumes. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(12):1322–1329. 1

[Cloyd et al., 2021] Cloyd, J., Haut, S., Carrazana, E., and Rabinowicz, A. L. (2021). Overcoming the challenges of developing an intranasal diazepam rescue therapy for the treatment of seizure clusters. *Epilepsia*, 62(4):846–856. xxix, 24

[Cook et al., 2013] Cook, M. J., O'Brien, T. J., Berkovic, S. F., Murphy, M., Morokoff, A., Fabinyi, G., D'Souza, W., Yerra, R., Archer, J., Litewka, L., et al. (2013). Prediction of Seizure Likelihood with a Long-Term, Implanted Seizure Advisory System in Patients with Drug-Resistant Epilepsy: a First-in-Man Study. *The Lancet Neurology*, 12(6):563–571. 2, 24, 26, 49, 50, 52, 53, 56, 57, 58, 59, 60, 61, 62, 67, 74, 77, 79, 85, 92, 123, 165, 173, 177, 185, 192, 223

[Cook, 1977] Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18. 42, 165

[Cormen et al., 2001] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). Introduction to Algorithms Second Edition. *The Knuth-Morris-Pratt Algorithm, year*. 88

[Cresswell et al., 2010] Cresswell, K. M., Worth, A., and Sheikh, A. (2010). Actor-network theory and its role in understanding the implementation of information technology developments in healthcare. *BMC medical informatics and decision making*, 10(1):67. 67, 68, 70, 71, 72

[Cui et al., 2018] Cui, S., Duan, L., Qiao, Y., and Xiao, Y. (2018). Learning eeg synchronization patterns for epileptic seizure prediction using bag-of-wave features. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–16. xxi, 26

[D'Alessandro et al., 2003] D'Alessandro, M., Esteller, R., Vachtsevanos, G., Hinson, A., Echauz, J., and Litt, B. (2003). Epileptic seizure prediction using hybrid feature selection over multiple intracranial eeg electrode contacts: a report of four patients. *IEEE transactions on biomedical engineering*, 50(5):603–615. 57

[Daoud and Bayoumi, 2019] Daoud, H. and Bayoumi, M. A. (2019). Efficient epileptic seizure prediction based on deep learning. *IEEE transactions on biomedical circuits and systems*, 13(5):804–813. 48, 50, 51, 52, 54, 55, 56, 57, 58, 60, 242

[Dash and Liu, 1997] Dash, M. and Liu, H. (1997). Feature Selection for Classification. *Intelligent data analysis*, 1(3):131–156. 224

[Davey and Adamopoulos, 2016] Davey, B. and Adamopoulos, A. (2016). Grounded theory and actor-network theory: a case study. *International Journal of Actor-Network Theory and Technological Innovation (IJANTTI)*, 8(1):27–33. 70

[De Tisi et al., 2011] De Tisi, J., Bell, G. S., Peacock, J. L., McEvoy, A. W., Harkness, W. F., Sander, J. W., and Duncan, J. S. (2011). The long-term outcome of adult epilepsy surgery, patterns of seizure remission, and relapse: a cohort study. *The Lancet*, 378(9800):1388–1395. 21

[Deb et al., 2002] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197. 87, 88, 209

[Debener et al., 2015] Debener, S., Emkes, R., De Vos, M., and Bleichner, M. (2015). Unobtrusive ambulatory eeg using a smartphone and flexible printed electrodes around the ear. *Scientific reports*, 5(1):1–11. 17, 166, 176, 193

[Dell et al., 2019] Dell, K. L., Cook, M. J., and Maturana, M. I. (2019). Deep brain stimulation for epilepsy: biomarkers for optimization. *Current Treatment Options in Neurology*, 21(10):1–16. 27

[Devinsky et al., 2016] Devinsky, O., Hesdorffer, D. C., Thurman, D. J., Lhatoo, S., and Richerson, G. (2016). Sudden unexpected death in epilepsy: epidemiology, mechanisms, and prevention. *The Lancet Neurology*, 15(10):1075–1088. 1

[Devinsky et al., 2018] Devinsky, O., Vezzani, A., O'Brien, T. J., Jette, N., Scheffer, I. E., de Curtis, M., and Perucca, P. (2018). Epilepsy. *Nature Reviews Disease Primers*, 4(1):18024. xxi, 11, 12, 13, 18, 19

[Direito et al., 2017] Direito, B., Teixeira, C. A., Sales, F., Castelo-Branco, M., and Dourado, A. (2017). A realistic seizure prediction study based on multiclass svm. *International journal of neural systems*, 27(03):1750006. 35, 36, 50, 51, 52, 54, 56, 58, 59, 61, 85, 87, 100, 104, 161, 162, 163, 223, 224

[Direito et al., 2011] Direito, B., Ventura, F., Teixeira, C., and Dourado, A. (2011). Optimized feature subsets for epileptic seizure prediction studies. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1636–1639. IEEE. 57, 224

[Dissanayake et al., 2021a] Dissanayake, T., Fernando, T., Denman, S., Sridharan, S., and Fookes, C. (2021a). Deep learning for patient-independent epileptic seizure prediction using scalp eeg signals. *IEEE Sensors Journal*, 21(7):9377–9388. 63, 64

[Dissanayake et al., 2021b] Dissanayake, T., Fernando, T., Denman, S., Sridharan, S., and Fookes, C. (2021b). Geometric deep learning for subject-independent epileptic seizure prediction using scalp eeg signals. *IEEE Journal of Biomedical and Health Informatics*. 63, 64

[Doshi-Velez and Kim, 2017] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. xxii, 2, 39, 40, 44, 80, 107, 123, 165, 167, 182

[Dreifuss et al., 1998] Dreifuss, F. E., Rosman, N. P., Cloyd, J. C., Pellock, J. M., Kuzniecky, R. I., Lo, W. D., Matsuo, F., Sharp, G. B., Conry, J. A., Bergen, D. C., et al. (1998). A comparison of rectal diazepam gel and placebo for acute repetitive seizures. *New England Journal of Medicine*, 338(26):1869–1875. 76, 100, 166

[Dumanis et al., 2017] Dumanis, S. B., French, J. A., Bernard, C., Worrell, G. A., and Fureman, B. E. (2017). Seizure forecasting from idea to reality. outcomes of the my seizure gauge epilepsy innovation institute workshop. *Eneuro*, 4(6). 27, 78, 116, 126

[Dümpelmann, 2019] Dümpelmann, M. (2019). Early seizure detection for closed loop direct neurostimulation devices in epilepsy. *Journal of neural engineering*, 16(4):041001. 61

[Duun-Henriksen et al., 2020] Duun-Henriksen, J., Baud, M., Richardson, M. P., Cook, M., Kouvas, G., Heasman, J. M., Friedman, D., Peltola, J., Zibrandtsen, I. C., and Kjaer, T. W. (2020). A new era in electroencephalographic monitoring? subscalp devices for ultra–long-term recordings. *Epilepsia*, 61(9):1805–1817. 17, 18, 125

[Edwards, 1960] Edwards, A. (1960). The meaning of binomial distribution. *Nature*, 186(4730):1074–1074. 256

[Eiben and Smith, 2003] Eiben, A. E. and Smith, J. E. (2003). *What is an Evolutionary Algorithm?*, pages 15–35. Springer Berlin Heidelberg, Berlin, Heidelberg. 84, 89, 215, 216

[Elger and Hoppe, 2018] Elger, C. E. and Hoppe, C. (2018). Diagnostic challenges in epilepsy: seizure under-reporting and seizure detection. *The Lancet Neurology*, 17(3):279–288. 11

[Engel, 2016] Engel, J. (2016). What can we do for people with drug-resistant epilepsy?: the 2016 wartenberg lecture. *Neurology*, 87(23):2483–2489. 2, 18, 19, 73, 166, 173, 174, 185

[Engel Jr, 2015] Engel Jr, J. (2015). Overview of surgical treatment for epilepsy. *The Treatment of Epilepsy*, pages 709–722. 19

[Engel Jr, 2018] Engel Jr, J. (2018). The current place of epilepsy surgery. *Current opinion in neurology*, 31(2):192. 19, 20, 21

[Ferastraoaru et al., 2018] Ferastraoaru, V., Goldenholz, D. M., Chiang, S., Moss, R., Theodore, W. H., and Haut, S. R. (2018). Characteristics of large patient-reported outcomes: where can one million seizures get us? *Epilepsia open*, 3(3):364–373. 62

[Feyissa et al., 2021] Feyissa, A. M., Worrell, G. A., and Lagerlund, T. D. (2021). EEG and Epilepsy. In Cascino, G. D., Sirven, J. I., and Tatum, W. O., editors, *Epilepsy*, chapter 6, pages 77–98. John Wiley & Sons, Ltd, 2nd edition. 14

[Fiest et al., 2017] Fiest, K. M., Sauro, K. M., Wiebe, S., Patten, S. B., Kwon, C.-S., Dykeman, J., Pringsheim, T., Lorenzetti, D. L., and Jetté, N. (2017). Prevalence and Incidence of Epilepsy: a Systematic Review and Meta-Analysis of International Studies. *Neurology*, 88(3):296–303. 1

[Fisher et al., 2014] Fisher, R. S., Acevedo, C., Arzimanoglou, A., Bogacz, A., Cross, J. H., Elger, C. E., Engel Jr, J., Forsgren, L., French, J. A., Glynn, M., et al. (2014). Ilae official report: a practical clinical definition of epilepsy. *Epilepsia*, 55(4):475–482. 10, 11

[Fisher et al., 2012] Fisher, R. S., Blum, D. E., DiVentura, B., Vannest, J., Hixson, J. D., Moss, R., Herman, S. T., Fureman, B. E., and French, J. A. (2012). Seizure diaries for clinical research and practice: limitations and future prospects. *Epilepsy & Behavior*, 24(3):304–310. 62

[Fisher et al., 2005] Fisher, R. S., Boas, W. V. E., Blume, W., Elger, C., Genton, P., Lee, P., and Engel Jr, J. (2005). Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe). *Epilepsia*, 46(4):470–472. 9, 10

[Fisher et al., 2017] Fisher, R. S., Cross, J. H., French, J. A., Higurashi, N., Hirsch, E., Jansen, F. E., Lagae, L., Moshé, S. L., Peltola, J., Roulet Perez, E., et al. (2017). Operational classification of seizure types by the international league against epilepsy: Position paper of the ilae commission for classification and terminology. *Epilepsia*, 58(4):522–530. xxi, 10, 12

[Forde and Paganini, 2019] Forde, J. Z. and Paganini, M. (2019). The scientific method in the science of machine learning. *arXiv preprint arXiv:1904.10922*. 121

[Foundation, 2018] Foundation, E. (2018). FDA Approval: Medtronic Deep Brain Stimulation for Medically Refractory Epilepsy. Epilepsy Foundation in https://www.epilepsy.com/article/2018/5/fda-approval-medtronic-deep-brain-stimulation-medically-refractory-epilepsy. 183

[Foundation, 2020] Foundation, E. (2020). Seizure Rescue Therapies: Oral Rescue Therapies. Epilepsy Foundation Website in https://www.epilepsy.com/learn/treating-seizures-and-epilepsy/seizure-rescue-therapies. 76, 166

[Freepik, 2021a] Freepik (2021a). Book Icon made by Freepik perfect from www.flaticon.com. FlaIcon Website in https://https://www.flaticon.com. xxiii, 69

[Freepik, 2021b] Freepik (2021b). Brainstorming Icon made by Freepik from www.flaticon.com. FlaIcon Website in https://https://www.flaticon.com. xxiii, 69

[Freestone et al., 2017] Freestone, D. R., Karoly, P. J., and Cook, M. J. (2017). A Forward-Looking Review of Seizure Prediction. *Current Opinion in Neurology*, 30(2):167–173. xxiv, 2, 3, 25, 37, 39, 48, 57, 59, 67, 69, 77, 79, 80, 81, 84, 101, 104, 122, 165, 167, 168, 173, 174, 176, 177, 178, 179, 181, 185, 191, 193

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232. 111, 165

[Gabeff et al., 2021] Gabeff, V., Teijeiro, T., Zapater, M., Cammoun, L., Rheims, S., Ryvlin, P., and Atienza, D. (2021). Interpreting deep learning models for epileptic seizure detection on eeg signals. *Artificial Intelligence in Medicine*, 117:102084. 63, 64

[Gadhoumi et al., 2013] Gadhoumi, K., Lina, J.-M., and Gotman, J. (2013). Seizure prediction in patients with mesial temporal lobe epilepsy using eeg measures of state similarity. *Clinical Neurophysiology*, 124(9):1745–1754. 163

[Gadhoumi et al., 2016a] Gadhoumi, K., Lina, J. M., Mormann, F., and Gotman, J. (2016a). Seizure Prediction for Therapeutic Devices: A Review. *Journal of Neuroscience Methods*, 260(029):270–282. 2, 25, 67, 78, 165, 178, 179, 180

[Gadhoumi et al., 2016b] Gadhoumi, K., Lina, J.-M., Mormann, F., and Gotman, J. (2016b). Seizure prediction for therapeutic devices: A review. *Journal of neuroscience methods*, 260:270–282. 24, 25, 29, 54, 67, 87

[Gagliano et al., 2019] Gagliano, L., Assi, E. B., Nguyen, D. K., and Sawan, M. (2019). Bispectrum and Recurrent neural Networks: Improved Classification of Interictal and Preictal States. *Scientific Reports*, 9(1):1–9. 84

[Gaínza-Lein et al., 2017] Gaínza-Lein, M., Benjamin, R., Stredny, C., McGurl, M., Kapur, K., and Loddenkemper, T. (2017). Rescue medications in epilepsy patients: a family perspective. *Seizure*, 52:188–194. 23, 78, 166, 183

[Gama et al., 2014] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37. xxii, 37

[Gayatri and Livingston, 2006] Gayatri, N. and Livingston, J. (2006). Aggravation of epilepsy by anti-epileptic drugs. *Developmental Medicine & Child Neurology*, 48(5):394–398. 13

[Gilpin et al., 2018] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE. 40, 41, 43, 44, 68, 107

[Gleaves et al., 2020] Gleaves, L. P., Schwartz, R., and Broniatowski, D. A. (2020). The role of individual user differences in interpretable and explainable machine learning systems. *arXiv preprint arXiv:2009.06675*. 40

[Goldstein et al., 2015] Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65. 111, 165, 182

[Goodman and Flaxman, 2017] Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57. 2, 39, 67, 122, 165, 167, 182, 185

[Gramfort et al., 2013] Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. S. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13. 98

[Grande et al., 2020] Grande, K. M., Ihnen, S. K., and Arya, R. (2020). Electrical stimulation mapping of brain function: a comparison of subdural electrodes and stereo-eeg. *Frontiers in Human Neuroscience*, page 538. 17

[Granger et al., 2003] Granger, C. B., Goldberg, R. J., Dabbous, O., Pieper, K. S., Eagle, K. A., Cannon, C. P., Van de Werf, F., Avezum, A., Goodman, S. G., Flather, M. D., et al. (2003). Predictors of hospital mortality in the global registry of acute coronary events. *Archives of internal medicine*, 163(19):2345–2353. 80, 122

[Grassberger, 1983] Grassberger, P. (1983). Generalized dimensions of strange attractors. *Physics Letters A*, 97(6):227–230. 163

[Grassberger and Procaccia, 1983] Grassberger, P. and Procaccia, I. (1983). Characterization of strange attractors. *Physical review letters*, 50(5):346. 163

[Greenwell et al., 2018] Greenwell, B. M., Boehmke, B. C., and McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*. 42

[Gregg et al., 2021] Gregg, N. M., Marks, V. S., Sladky, V., Lundstrom, B. N., Klassen, B., Messina, S. A., Brinkmann, B. H., Miller, K. J., Van Gompel, J. J., Kremen, V., et al. (2021). Anterior nucleus of the thalamus seizure detection in ambulatory humans. *Epilepsia*, 62(10):e158–e164. 125

[Harmer et al., 2017] Harmer, C. J., Duman, R. S., and Cowen, P. J. (2017). How do antidepressants work? new perspectives for refining future treatment approaches. *The Lancet Psychiatry*, 4(5):409–418. 81

[Harrison et al., 2005] Harrison, M. A. F., Osorio, I., Frei, M. G., Asuri, S., and Lai, Y.-C. (2005). Correlation dimension and integral do not predict epileptic seizures. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 15(3):033106. 54

[Health, 2017] Health, A. K. (2017). Invasive electroencephalography (EEG) monitoring before epilepsy surgery. 17

[Herman, 2017] Herman, B. (2017). The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414*. 43

[Hoens et al., 2012] Hoens, T. R., Polikar, R., and Chawla, N. V. (2012). Learning from streaming data with concept drift and imbalance: an overview. *Progress in Artificial Intelligence*, 1(1):89–101. 36

[Hossain et al., 2019] Hossain, M. S., Amin, S. U., Alsulaiman, M., and Muhammad, G. (2019). Applying deep learning for epilepsy seizure detection and brain mapping visualization. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1s):1–17. 63, 64

[Iasemidis, 2003] Iasemidis, L. D. (2003). Epileptic seizure prediction and control. *IEEE Transactions on Biomedical Engineering*, 50(5):549–558. 9, 14, 67, 163

[Iasemidis et al., 2004] Iasemidis, L. D., Shiau, D.-S., Sackellares, J. C., Pardalos, P. M., and Prasad, A. (2004). Dynamical resetting of the human brain at epileptic seizures: application of nonlinear dynamics and global optimization techniques. *IEEE transactions on biomedical engineering*, 51(3):493–506. 164

[Icon, 2021] Icon, F. (2021). Guidelines Icon made by Flat Icon from www.flaticon.com. FlaIcon Website in https://https://www.flaticon.com. xxiii, 69

[Ihle et al., 2012] Ihle, M., Feldwisch-Drentrup, H., Teixeira, C. A., Witon, A., Schelter, B., Timmer, J., and Schulze-Bonhage, A. (2012). EPILEPSIAE - A European Epilepsy Database. *Comput. Methods Programs Biomed.*, 106(3):127–138. 1, 85, 108

[Inglis et al., 2022] Inglis, A., Parnell, A., and Hurley, C. B. (2022). Visualizing variable importance and variable interaction effects in machine learning models. *Journal of Computational and Graphical Statistics*, pages 1–13. 42

[Islam et al., 2020] Islam, M. S., El-Hajj, A. M., Alawieh, H., Dawy, Z., Abbas, N., and El-Imad, J. (2020). Eeg mobility artifact removal for ambulatory epileptic seizure prediction applications. *Biomedical Signal Processing and Control*, 55:101638. 166

[Iyamu and Mgudlwa, 2018] Iyamu, T. and Mgudlwa, S. (2018). Transformation of healthcare big data through the lens of actor network theory. *International Journal of Healthcare Management*, 11(3):182–192. 71

[Jacobs et al., 2018] Jacobs, D., Hilton, T., Del Campo, M., Carlen, P. L., and Bardakjian, B. L. (2018). Classification of pre-clinical seizure states using scalp eeg cross-frequency coupling features. *IEEE Transactions on Biomedical Engineering*, 65(11):2440–2449. 87

[Jafarpour et al., 2019] Jafarpour, S., Hirsch, L. J., Gaínza-Lein, M., Kellinghaus, C., and Detyniecki, K. (2019). Seizure cluster: definition, prevalence, consequences, and management. *Seizure*, 68:9–15. 14, 166, 176, 178

[Jayakar et al., 2016] Jayakar, P., Gotman, J., Harvey, A. S., Palmini, A., Tassi, L., Schomer, D., Dubeau, F., Bartolomei, F., Yu, A., Kršek, P., et al. (2016). Diagnostic utility of invasive eeg for epilepsy surgery: indications, modalities, and techniques. *Epilepsia*, 57(11):1735–1747. 17

[Jette and Engel, 2016] Jette, N. and Engel, J. (2016). Refractory Epilepsy is a Life-Threatening Disease: Lest we Forget. 1

[Jia and Kohn, 2011] Jia, X. and Kohn, A. (2011). Gamma Rhythms in the Brain. *PLoS Biology*, 9(4). 87

[Jones and Thomas, 2017] Jones, L. A. and Thomas, R. H. (2017). Sudden death in epilepsy: Insights from the last 25 years. *Seizure*, 44:232–236. 1

[Jurcak et al., 2007] Jurcak, V., Tsuzuki, D., and Dan, I. (2007). 10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems. *Neuroimage*, 34(4):1600–1611. 16

[Kaminski, 2019] Kaminski, M. E. (2019). The right to explanation, explained. *Berkeley Tech. LJ*, 34:189. 122

[Karoly et al., 2020] Karoly, P. J., Cook, M. J., Maturana, M., Nurse, E. S., Payne, D., Brinkmann, B. H., Grayden, D. B., Dumanis, S. B., Richardson, M. P., Worrell, G. A., et al. (2020). Forecasting cycles of seizure likelihood. *Epilepsia*, 61(4):776–786. 122, 123, 126

[Karoly et al., 2016] Karoly, P. J., Freestone, D. R., Boston, R., Grayden, D. B., Himes, D., Leyde, K., Seneviratne, U., Berkovic, S., O'Brien, T., and Cook, M. J. (2016). Interictal spikes and epileptic seizures: their relationship and underlying rhythmicity. *Brain*, 139(4):1066–1078. 37

[Karoly et al., 2018] Karoly, P. J., Goldenholz, D. M., Freestone, D. R., Moss, R. E., Grayden, D. B., Theodore, W. H., and Cook, M. J. (2018). Circadian and circaseptan rhythms in human epilepsy: a retrospective cohort study. *The Lancet Neurology*, 17(11):977–985. 62

[Karoly et al., 2021] Karoly, P. J., Rao, V. R., Gregg, N. M., Worrell, G. A., Bernard, C., Cook, M. J., and Baud, M. O. (2021). Cycles in epilepsy. *Nature Reviews Neurology*, 17(5):267–284. 61, 62

[Karoly et al., 2017] Karoly, P. J., Ung, H., Grayden, D. B., Kuhlmann, L., Leyde, K., Cook, M. J., and Freestone, D. R. (2017). The circadian profile of epilepsy improves seizure forecasting. *Brain*, 140(8):2169–2182. 2, 26, 50, 52, 54, 56, 57, 58, 60, 61, 62, 122, 123, 165

[Khan et al., 2017] Khan, H., Marcuse, L., Fields, M., Swann, K., and Yener, B. (2017). Focal onset seizure prediction using convolutional networks. *IEEE Transactions on Biomedical Engineering*, 65(9):2109–2118. 60, 84

[Khan et al., 2018] Khan, S., Nobili, L., Khatami, R., Loddenkemper, T., Cajochen, C., Dijk, D.-J., and Eriksson, S. H. (2018). Circadian rhythm and epilepsy. *The Lancet Neurology*, 17(12):1098–1108. 37, 62

[Kim et al., 2016] Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29. 40, 165, 182

[Kiral-Kornek et al., 2018] Kiral-Kornek, I., Roy, S., Nurse, E., Mashford, B., Karoly, P., Carroll, T., Payne, D., Saha, S., Baldassano, S., O'Brien, T., et al. (2018). Epileptic seizure prediction using big data and deep learning: toward a mobile system. *EBioMedicine*, 27:103–111. 49, 50, 52, 53, 54, 56, 58, 59, 61

[Kirby et al., 2020] Kirby, J., Leach, V. M., Brockington, A., Patsalos, P., Reuber, M., and Leach, J. P. (2020). Drug withdrawal in the epilepsy monitoring unit–the patsalos table. *Seizure*, 75:75–81. 19

[Klatt et al., 2012] Klatt, J., Feldwisch-Drentrup, H., Ihle, M., Navarro, V., Neufang, M., Teixeira, C., Adam, C., Valderrama, M., Alvarado-Rojas, C., Witon, A., Le Van Quyen, M., Sales, F., Dourado, A., Timmer, J., Schulze-Bonhage, A., and Schelter, B. (2012). The EPILEPSIAE Database: An Extensive Electroencephalography Database of Epilepsy Patients. *Epilepsia*, 53(9):1669–1676. 1, 3, 13, 85, 108

[Kodankandath et al., 2020] Kodankandath, T. V., Theodore, D., and Samanta, D. (2020). Generalized tonic-clonic seizure. *StatPearls*. 12

[Koh and Liang, 2017] Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR. 42

[Kramer and Cash, 2012] Kramer, M. A. and Cash, S. S. (2012). Epilepsy as a disorder of cortical network organization. *The Neuroscientist*, 18(4):360–372. 87

[Kramer et al., 2010] Kramer, M. A., Eden, U. T., Kolaczyk, E. D., Zepeda, R., Eskandar, E. N., and Cash, S. S. (2010). Coalescence and fragmentation of cortical networks during focal seizures. *Journal of Neuroscience*, 30(30):10076–10085. 87

[Kramer, 2016] Kramer, O. (2016). Scikit-learn. In *Machine learning for evolution strategies*, pages 45–53. Springer. 109, 110

[Kreuz et al., 2004] Kreuz, T., Andrzejak, R. G., Mormann, F., Kraskov, A., Stögbauer, H., Elger, C. E., Lehnertz, K., and Grassberger, P. (2004). Measure profile surrogates: a method to validate the performance of epileptic seizure prediction algorithms. *Physical Review E*, 69(6):061915. 34

[Krishna et al., 2016] Krishna, V., Sammartino, F., King, N. K. K., So, R. Q. Y., and Wennberg, R. (2016). Neuromodulation for epilepsy. *Neurosurgery Clinics*, 27(1):123–131. 21

[Kuhlmann et al., 2018a] Kuhlmann, L., Karoly, P., Freestone, D. R., Brinkmann, B. H., Temko, A., Barachant, A., Li, F., Titericz Jr, G., Lang, B. W., Lavery, D., et al. (2018a). Epilepsyecosystem. org: crowd-sourcing reproducible seizure prediction with long-term human intracranial eeg. *Brain*, 141(9):2619–2630. 49, 50, 52, 54, 56, 57, 58, 104, 161, 162, 163, 164

[Kuhlmann et al., 2018b] Kuhlmann, L., Lehnertz, K., Richardson, M. P., Schelter, B., and Zaveri, H. P. (2018b). Seizure Prediction — Ready for a New Era. *Nature Reviews Neurology*, 14(10):618–630. xxiv, 2, 3, 25, 29, 37, 39, 48, 57, 59, 67, 69, 76, 81, 84, 101, 105, 122, 126, 165, 167, 169, 173, 174, 176, 178, 179, 180, 181, 184, 185, 193

[Kwan et al., 2010] Kwan, P., Arzimanoglou, A., Berg, A. T., Brodie, M. J., Allen Hauser, W., Mathern, G., Moshé, S. L., Perucca, E., Wiebe, S., and French, J. (2010). Definition of drug resistant epilepsy: consensus proposal by the ad hoc task force of the ilae commission on therapeutic strategies. 18, 19

[Lage et al., 2019] Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., and Doshi-Velez, F. (2019). An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*. 2, 165, 182

[Laxer et al., 2014] Laxer, K. D., Trinka, E., Hirsch, L. J., Cendes, F., Langfitt, J., Delanty, N., Resnick, T., and Benbadis, S. R. (2014). The Consequences of Refractory Epilepsy and its Treatment. *Epilepsy & Behavior*, 37:59–70. 1, 18, 21, 22, 23

[Le Van Quyen et al., 1999] Le Van Quyen, M., Martinerie, J., Baulac, M., and Varela, F. (1999). Anticipating epileptic seizures in real time by a non-linear analysis of similarity between eeg recordings. *Neuroreport*, 10(10):2149–2155. 163

[Leal et al., 2021] Leal, A., Pinto, M. F., Lopes, F., Bianchi, A. M., Henriques, J., Ruano, M. G., de Carvalho, P., Dourado, A., and Teixeira, C. A. (2021). Heart rate variability analysis for the identification of the preictal interval in patients with drug-resistant epilepsy. *Scientific reports*, 11(1):1–11. 54

[Lee et al., 2020] Lee, T., Hwang, J., and Lee, H. (2020). Trier: Template-guided neural networks for robust and interpretable sleep stage identification from eeg recordings. *arXiv preprint arXiv:2009.05407*. 63, 64

[Leguia et al., 2021] Leguia, M. G., Andrzejak, R. G., Rummel, C., Fan, J. M., Mirro, E. A., Tcheng, T. K., Rao, V. R., and Baud, M. O. (2021). Seizure cycles in focal epilepsy. *JAMA neurology*, 78(4):454–463. 62

[Lewis, 2019] Lewis, D. (2019). History and perspective on diy closed looping. *Journal of diabetes science and technology*, 13(4):790–793. 81, 166

[Lin et al., 2019] Lin, Z. Q., Shafiee, M. J., Bochkarev, S., Jules, M. S., Wang, X. Y., and Wong, A. (2019). Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms. *arXiv preprint arXiv:1910.07387*. 42

[Litt et al., 2001] Litt, B., Esteller, R., Echauz, J., D'Alessandro, M., Shor, R., Henry, T., Pennell, P., Epstein, C., Bakay, R., Dichter, M., et al. (2001). Epileptic seizures may begin hours in advance of clinical onset: a report of five patients. *Neuron*, 30(1):51–64. 115, 118, 119

[Liu and Motoda, 1998] Liu, H. and Motoda, H. (1998). *Feature Extraction, Construction and Selection: A Data Mining Perspective*, volume 453. Springer Science & Business Media. 85

[Lo Giudice et al., 2022] Lo Giudice, M., Varone, G., Ieracitano, C., Mammone, N., Tripodi, G. G., Ferlazzo, E., Gasparini, S., Aguglia, U., and Morabito, F. C. (2022). Permutation entropy-based interpretability of convolutional neural network models for interictal eeg discrimination of subjects with epileptic seizures vs. psychogenic non-epileptic seizures. *Entropy*, 24(1):102. 63, 64

[Lombrozo, 2006] Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470. 77, 165, 181

[Lopes et al., 2021] Lopes, F., Leal, A., Medeiros, J., Pinto, M. F., Dourado, A., Dümpelmann, M., and Teixeira, C. (2021). Automatic electroencephalogram artifact removal using deep convolutional neural networks. *IEEE Access*, 9:149955–149970. 26, 109

[Lu et al., 2018] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363. xxii, 36, 37

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. 111, 165, 182

[Maina et al., 2021] Maina, M. B., Ahmad, U., Ibrahim, H. A., Hamidu, S., Nasr, F. E., Salihu, A., Abushouk, A. I., Abdurrazak, M., Awadelkareem, M., Amin, A., et al. (2021). Two decades of neuroscience publication trends in africa. *Nature communications*, 12(1):1–10. 126

[Majety et al., 2021] Majety, R. P. D., Katru, S., Veluchuri, J. P., and Juturi, R. K. R. (2021). New era in medical device regulations in the european union. *Pharmaceut Reg Affairs*, 10:234. 2, 166

[Mansour et al., 2020] Mansour, M., Khnaisser, F., and Partamian, H. (2020). An explainable model for eeg seizure detection based on connectivity features. *arXiv preprint arXiv:2009.12566*. 63, 64

[Markert and Fisher, 2019] Markert, M. S. and Fisher, R. S. (2019). Neuromodulation-science and practice in epilepsy: vagus nerve stimulation, thalamic deep brain stimulation, and responsive neurostimulation. *Expert review of neurotherapeutics*, 19(1):17–29. 22, 23

[Maturana et al., 2020] Maturana, M. I., Meisel, C., Dell, K., Karoly, P. J., D'Souza, W., Grayden, D. B., Burkitt, A. N., Jiruska, P., Kudlacek, J., Hlinka, J., et al. (2020). Critical slowing down as a biomarker for seizure susceptibility. *Nature communications*, 11(1):1–12. 27

[McLane et al., 2015] McLane, H. C., Berkowitz, A. L., Patenaude, B. N., McKenzie, E. D., Wolper, E., Wahlster, S., Fink, G., and Mateen, F. J. (2015). Availability, accessibility, and affordability of neurodiagnostic tests in 37 countries. *Neurology*, 85(18):1614–1622. 126

[McSharry et al., 2003] McSharry, P. E., Smith, L. A., and Tarassenko, L. (2003). Prediction of epileptic seizures: are nonlinear methods relevant? *Nature medicine*, 9(3):241–242. 54

[Medithe and Nelakuditi, 2016] Medithe, J. W. C. and Nelakuditi, U. R. (2016). Study of normal and abnormal eeg. In *2016 3rd International conference on advanced computing and communication systems (ICACCS)*, volume 1, pages 1–4. IEEE. 14, 16

[Meisel and Bailey, 2019] Meisel, C. and Bailey, K. A. (2019). Identifying signal-dependent information about the preictal state: a comparison across ecog, eeg and ekg using deep learning. *EBioMedicine*, 45:422–431. 26

[Meisel et al., 2020] Meisel, C., El Atrache, R., Jackson, M., Schubach, S., Ufongene, C., and Loddenkemper, T. (2020). Machine learning from wristband sensor data for wearable, noninvasive seizure forecasting. *Epilepsia*, 61(12):2653–2666. 26

[Miller, 2019] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38. 41, 107

[Miller et al., 2017] Miller, T., Howe, P., and Sonenberg, L. (2017). Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*. 39, 41, 107

[Mirowski et al., 2008] Mirowski, P. W., LeCun, Y., Madhavan, D., and Kuzniecky, R. (2008). Comparing SVM and Convolutional Networks for Epileptic Seizure Prediction from Intracranial EEG. In *2008 IEEE Workshop on Machine Learning for Signal Processing*, pages 244–249. IEEE. 84

[Mitchell and Taylor, 1999] Mitchell, M. and Taylor, C. E. (1999). Evolutionary computation: an overview. *Annual Review of Ecology and Systematics*, 30(1):593–616. 84

[Moghaddam et al., 2022] Moghaddam, D. E., Sheth, S., Haneef, Z., Gavvala, J., and Aazhang, B. (2022). Epileptic seizure prediction using spectral width of the covariance matrix. *Journal of Neural Engineering*. 63, 64

[Moghim and Corne, 2014] Moghim, N. and Corne, D. W. (2014). Predicting epileptic seizures in advance. *PloS one*, 9(6):e99334. 50, 52, 53, 56, 57, 58, 59, 84, 161, 162, 163

[Mohseni et al., 2006] Mohseni, H. R., Maghsoudi, A., and Shamsollahi, M. B. (2006). Seizure detection in eeg signals: A comparison of different approaches. In *2006 International conference of the IEEE engineering in medicine and biology society*, pages 6724–6727. IEEE. 40

[Molnar, 2019] Molnar, C. (2019). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/. xxiv, 2, 39, 40, 41, 42, 44, 69, 77, 80, 91, 107, 111, 165, 167, 171, 181, 182, 192

[Mormann et al., 2007] Mormann, F., Andrzejak, R. G., Elger, C. E., and Lehnertz, K. (2007). Seizure Prediction: The Long and Winding Road. *Brain*, 130(2):314–333. xxiv, 2, 3, 14, 25, 26, 27, 28, 29, 30, 34, 54, 67, 69, 73, 79, 80, 81, 83, 92, 101, 103, 109, 126, 161, 163, 165, 168, 173, 174, 175, 176, 177, 178, 179, 180, 181, 183, 185, 193

[Mormann et al., 2003] Mormann, F., Kreuz, T., Andrzejak, R. G., David, P., Lehnertz, K., and Elger, C. E. (2003). Epileptic seizures are preceded by a decrease in synchronization. *Epilepsy research*, 53(3):173–185. 164

[Mormann et al., 2005] Mormann, F., Kreuz, T., Rieke, C., Andrzejak, R. G., Kraskov, A., David, P., Elger, C. E., and Lehnertz, K. (2005). On the predictability of epileptic seizures. *Clinical neurophysiology*, 116(3):569–587. 54, 161, 162, 163, 164

[Motoda and Liu, 2002] Motoda, H. and Liu, H. (2002). Feature Selection, Extraction and Construction. *Communication of IICM (Institute of Information and Computing Machinery, Taiwan) Vol*, 5(67-72):2. 85

[Mporas et al., 2015] Mporas, I., Tsirka, V., Zacharaki, E. I., Koutroumanidis, M., Richardson, M., and Megalooikonomou, V. (2015). Seizure detection using eeg and ecg signals for computer-based monitoring, analysis and management of epileptic patients. *Expert systems with applications*, 42(6):3227–3233. 14

[Müller et al., 2022] Müller, J., Yang, H., Eberlein, M., Leonhardt, G., Uckermann, O., Kuhlmann, L., and Tetzlaff, R. (2022). Coherent false seizure prediction in epilepsy, coincidence or providence? *Clinical Neurophysiology*, 133:157–164. 54

[Murphy et al., 2010] Murphy, M., Smith, P. D., Wood, M., Bowden, S., O'Brien, T. J., Bulluss, K. J., and Cook, M. J. (2010). Surgery for temporal lobe epilepsy associated with mesial temporal sclerosis in the older patient: a long-term follow-up. *Epilepsia*, 51(6):1024–1029. 21

[Nasseri et al., 2020] Nasseri, M., Nurse, E., Glasstetter, M., Böttcher, S., Gregg, N. M., Laks Nandakumar, A., Joseph, B., Pal Attia, T., Viana, P. F., Bruno, E., et al. (2020). Signal quality and patient experience with wearable devices for epilepsy management. *Epilepsia*, 61:S25–S35. 17, 24, 54, 125

[Nasseri et al., 2021] Nasseri, M., Pal Attia, T., Joseph, B., Gregg, N. M., Nurse, E. S., Viana, P. F., Worrell, G., Dümpelmann, M., Richardson, M. P., Freestone, D. R., et al. (2021). Ambulatory seizure forecasting with a wrist-worn device using long-short term memory deep learning. *Scientific reports*, 11(1):1–9. 26, 48, 50, 52, 56, 57, 58, 59, 60, 61

[Naze et al., 2021] Naze, S., Tang, J., Kozloski, J. R., and Harrer, S. (2021). Features importance in seizure classification using scalp eeg reduced to single time-series. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 329–332. IEEE. 63, 64

[Noachtar and Rémi, 2009] Noachtar, S. and Rémi, J. (2009). The role of eeg in epilepsy: a critical review. *Epilepsy & Behavior*, 15(1):22–33. 17, 18

[Nunez and Srinivasan, 2006] Nunez, P. L. and Srinivasan, R. (2006). Recording strategies, reference issues, and dipole localization. *Nunez PL, Srinivasan R: Electric Fields of the Brain: The Neurophysics of EEG, ed*, 2:275–312. 16

[Nurse et al., 2016] Nurse, E., Mashford, B. S., Yepes, A. J., Kiral-Kornek, I., Harrer, S., and Freestone, D. R. (2016). Decoding eeg and lfp signals using deep learning: heading truenorth. In *Proceedings of the ACM international conference on computing frontiers*, pages 259–266. 80, 165, 184

[Oja and Hyvarinen, 2000] Oja, E. and Hyvarinen, A. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430. 164

[Okamoto et al., 2004] Okamoto, M., Dan, H., Sakamoto, K., Takeo, K., Shimizu, K., Kohno, S., Oda, I., Isobe, S., Suzuki, T., Kohyama, K., et al. (2004). Three-dimensional probabilistic anatomical cranio-cerebral correlation via the international 10–20 system oriented for transcranial functional brain mapping. *Neuroimage*, 21(1):99–111. 98

[Okun and Foote, 2010] Okun, M. S. and Foote, K. D. (2010). Parkinson's disease dbs: what, when, who and why? the time has come to tailor dbs targets. *Expert review of neurotherapeutics*, 10(12):1847–1857. 80

[Oliveira, 2021] Oliveira, A. C. R. (2021). Sleep-awake cycle evaluation from long-term eeg data: Assessing the impact in epilepsy seizure prediction. Master's thesis, Universidade de Coimbra. 111

[Oliveira et al., 2016] Oliveira, A. S., Schlink, B. R., Hairston, W. D., König, P., and Ferris, D. P. (2016). Induction and separation of motion artifacts in eeg data using a mobile phantom head device. *Journal of neural engineering*, 13(3):036014. 15

[Osorio et al., 2016] Osorio, I., Zaveri, H. P., Frei, M. G., and Arthurs, S. (2016). *Epilepsy: the intersection of neurosciences, biology, mathematics, engineering, and physics*. CRC press. 17, 25, 26

[Owolabi et al., 2018] Owolabi, L. F., Sale, S., Owolabi, S. D., Nalado, A., Umar, M., and Taura, A. A. (2018). Electroencephalography abnormalities in generalized epilepsy and their predictors: A multicenter experience. *Annals of African medicine*, 17(2):64. 16

[Pal Attia et al., 2022] Pal Attia, T., Viana, P. F., Nasseri, M., Duun-Henriksen, J., Biondi, A., Winston, J. S., Martins, I. P., Nurse, E. S., Dümpelmann, M., Worrell, G. A., et al. (2022). Seizure forecasting using minimally-invasive, ultra long-term subcutaneous eeg: Generalizable cross-patient models. *Epilepsia*. 48, 50, 52, 54, 56, 57, 58, 60

[Park et al., 2011] Park, Y., Luo, L., Parhi, K. K., and Netoff, T. (2011). Seizure prediction with spectral power of eeg using cost-sensitive support vector machines. *Epilepsia*, 52(10):1761–1770. 34, 51, 53, 54, 58, 84

[Phan et al., 2022] Phan, H., Mikkelsen, K. B., Chen, O., Koch, P., Mertins, A., and De Vos, M. (2022). Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*. 63

[Pinto et al., 2022] Pinto, M., Coelho, T., Leal, A., Lopes, F., Dourado, A., Martins, P., and Teixeira, C. (2022). Interpretable eeg seizure prediction using a multiobjective evolutionary algorithm. *Scientific reports*, 12(1):1–15. 83, 101, 109, 115, 121, 256

[Pinto et al., 2021a] Pinto, M., Leal, A., Lopes, F., Dourado, A., Martins, P., Teixeira, C. A., et al. (2021a). A personalized and evolutionary algorithm for interpretable eeg epilepsy seizure prediction. *Scientific reports*, 11(1):1–12. 83, 101, 109, 121, 256

[Pinto et al., 2021b] Pinto, M. F., Leal, A., Lopes, F., Pais, J., Dourado, A., Sales, F., Martins, P., and Teixeira, C. A. (2021b). On the clinical acceptance of black-box systems for eeg seizure prediction. *Epilepsia Open*. 67, 122

[Priyasad et al., 2021] Priyasad, D., Fernando, T., Denman, S., Sridharan, S., and Fookes, C. (2021). Interpretable seizure classification using unprocessed eeg with multi-channel attentive feature fusion. *IEEE Sensors Journal*, 21(17):19186–19197. 63, 64

[Proix et al., 2021] Proix, T., Truccolo, W., Leguia, M. G., Tcheng, T. K., King-Stephens, D., Rao, V. R., and Baud, M. O. (2021). Forecasting seizure risk in adults with focal epilepsy: a development and validation study. *The Lancet Neurology*, 20(2):127–135. 27, 126

[Rabbi et al., 2013] Rabbi, A. F., Azinfar, L., and Fazel-Rezai, R. (2013). Seizure prediction using adaptive neuro-fuzzy inference system. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2100–2103. IEEE. 50, 51, 52, 53, 54, 56, 58, 59, 61, 163, 164

[Ramachandran et al., 2017] Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*. 243

[Ramgopal et al., 2014] Ramgopal, S., Thome-Souza, S., Jackson, M., Kadish, N. E., Fernández, I. S., Klehm, J., Bosl, W., Reinsberger, C., Schachter, S., and Loddenkemper, T. (2014). Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. *Epilepsy & behavior*, 37:291–307. xxiv, 24, 27, 67, 69, 165, 167, 170, 175, 183, 184

[Rao et al., 2021] Rao, V. R., G Leguia, M., Tcheng, T. K., and Baud, M. O. (2021). Cues for seizure timing. *Epilepsia*, 62:S15–S31. 61, 62

[Rasekhi et al., 2013] Rasekhi, J., Mollaei, M. R. K., Bandarabadi, M., Teixeira, C. A., and Dourado, A. (2013). Preprocessing effects of 22 linear univariate features on the performance of seizure prediction methods. *Journal of neuroscience methods*, 217(1-2):9–16. 50, 51, 52, 53, 54, 56, 58, 59, 87, 100, 161, 162, 163

[Rasekhi et al., 2015] Rasekhi, J., Mollaei, M. R. K., Bandarabadi, M., Teixeira, C. A., and Dourado, A. (2015). Epileptic seizure prediction based on ratio and differential linear univariate features. *Journal of medical signals and sensors*, 5(1):1. 50, 51, 52, 54, 56, 57, 58, 59, 61, 161, 162

[Rathore and Radhakrishnan, 2015] Rathore, C. and Radhakrishnan, K. (2015). Concept of epilepsy surgery and presurgical evaluation. *Epileptic disorders*, 17(1):19–31. 19, 20, 166, 174

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. 42, 81, 111, 165, 182

[Ribeiro et al., 2018] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). 165, 182

[Rincon et al., 2021] Rincon, N., Barr, D., and Velez-Ruiz, N. (2021). Neuromodulation in drug resistant epilepsy. *Aging and disease*, 12(4):1070. 21, 22, 23

[Riss et al., 2008] Riss, J., Cloyd, J., Gates, J., and Collins, S. (2008). Benzodiazepines in epilepsy: pharmacology and pharmacokinetics. *Acta neurologica scandinavica*, 118(2):69–86. 23

[Rosenberg and Van Hout, 2013] Rosenberg, R. S. and Van Hout, S. (2013). The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring. *Journal of clinical sleep medicine*, 9(1):81–87. 38

[Rubboli and Gardella, 2019] Rubboli, G. and Gardella, E. (2019). Non-age-related focal epilepsies. In *Clinical Electroencephalography*, pages 445–460. Springer. 108

[Rudin, 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215. 68, 84, 122

[Ryvlin et al., 2014] Ryvlin, P., Cross, J. H., and Rheims, S. (2014). Epilepsy surgery in children and adults. *The Lancet Neurology*, 13(11):1114–1126. 19, 21

[Ryvlin et al., 2021] Ryvlin, P., Rheims, S., Hirsch, L. J., Sokolov, A., and Jehi, L. (2021). Neuromodulation in epilepsy: state-of-the-art approved therapies. *The Lancet Neurology*, 20(12):1038–1047. xxi, xxix, 21, 22, 23

[Sanei and Chambers, 2007] Sanei, S. and Chambers, J. (2007). Introduction to eeg. *EEG signal processing*, pages 1–34. xxi, 14, 15

[Scheepers et al., 2000] Scheepers, M., Scheepers, B., Clarke, M., Comish, S., and Ibitoye, M. (2000). Is intranasal midazolam an effective rescue medication in adolescents and adults with severe epilepsy? *Seizure*, 9(6):417–421. 78, 166, 183

[Scheffer et al., 2017] Scheffer, I. E., Berkovic, S., Capovilla, G., Connolly, M. B., French, J., Guilhoto, L., Hirsch, E., Jain, S., Mathern, G. W., Moshé, S. L., et al. (2017). Ilae classification of the epilepsies: position paper of the ilae commission for classification and terminology. *Epilepsia*, 58(4):512–521. xxi, 10, 11, 13, 166, 176

[Scheffer et al., 2009] Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., Van Nes, E. H., Rietkerk, M., and Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, 461(7260):53–59. 115, 119

[Schelter et al., 2008] Schelter, B., Andrzejak, R. G., and Mormann, F. (2008). Can your prediction algorithm beat a random predictor? *Seizure prediction in epilepsy: from basic mechanisms to clinical applications*, pages 237–248. xxii, 30, 31, 32, 34, 35

[Schelter et al., 2007] Schelter, B., Winterhalder, M., genannt Drentrup, H. F., Wohlmuth, J., Nawrath, J., Brandt, A., Schulze-Bonhage, A., and Timmer, J. (2007). Seizure prediction: the impact of long prediction horizons. *Epilepsy research*, 73(2):213–217. 29

[Schelter et al., 2006] Schelter, B., Winterhalder, M., Maiwald, T., Brandt, A., Schad, A., Timmer, J., and Schulze-Bonhage, A. (2006). Do false predictions of seizures depend on the state of vigilance? a report from two seizure-prediction methods and proposed remedies. *Epilepsia*, 47(12):2058–2070. 62, 100, 101, 165, 180

[Schirrmeister et al., 2017] Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., and Ball, T. (2017). Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420. 60, 63, 64, 80, 84, 166

[Scholkopf et al., 1997] Scholkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. (1997). Comparing support vector machines with

gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45(11):2758–2765. 59

[Schulze-Bonhage, 2019] Schulze-Bonhage, A. (2019). Long-term outcome in neurostimulation of epilepsy. *Epilepsy & Behavior*, 91:25–29. xxix, 22, 23

[Schulze-Bonhage et al., 2010] Schulze-Bonhage, A., Sales, F., Wagner, K., Teotonio, R., Carius, A., Schelle, A., and Ihle, M. (2010). Views of patients with epilepsy on seizure prediction devices. *Epilepsy & behavior*, 18(4):388–396. 30, 67, 69, 79, 100, 126, 165, 192, 193, 197

[Scott, 1988] Scott, J. (1988). Social network analysis. *Sociology*, 22(1):109–127. 68, 70

[Shahar, 1997] Shahar, E. (1997). A popperian perspective of the term 'evidence-based medicine'. *Journal of Evaluation in Clinical practice*, 3(2):109–116. 121

[Shrikumar et al., 2017] Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR. 42

[Siddiqui et al., 2020] Siddiqui, M. K., Morales-Menendez, R., Huang, X., and Hussain, N. (2020). A review of epileptic seizure detection using machine learning classifiers. *Brain informatics*, 7(1):1–18. 40

[Simonyan et al., 2013] Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*. 42

[Sisterson et al., 2020] Sisterson, N. D., Wozny, T. A., Kokkinos, V., Bagic, A., Urban, A. P., and Richardson, R. M. (2020). A rational approach to understanding and evaluating responsive neurostimulation. *Neuroinformatics*, pages 1–11. 80, 166, 194

[Sondhi, 2009] Sondhi, P. (2009). Feature Construction Methods: a Survey. *Sifaka. Cs. Uiuc. Edu*, 69:70–71. 85

[Spencer et al., 2015] Spencer, D., Nguyen, D. K., and Sivaraju, A. (2015). Invasive eeg in presurgical evaluation of epilepsy. *The treatment of epilepsy*, pages 733–755. 17, 18

[Stacey et al., 2020] Stacey, W., Kramer, M., Gunnarsdottir, K., Gonzalez-Martinez, J., Zaghloul, K., Inati, S., Sarma, S., Stiso, J., Khambhati, A. N., Bassett, D. S., et al. (2020). Emerging roles of network analysis for epilepsy. *Epilepsy research*, 159:106255. 87

[Stirling et al., 2021a] Stirling, R. E., Cook, M. J., Grayden, D. B., and Karoly, P. J. (2021a). Seizure forecasting and cyclic control of seizures. *Epilepsia*, 62:S2–S14. 122, 123

[Stirling et al., 2021b] Stirling, R. E., Grayden, D. B., D'Souza, W., Cook, M. J., Nurse, E., Freestone, D. R., Payne, D. E., Brinkmann, B. H., Pal Attia, T., Viana, P. F., et al. (2021b). Forecasting seizure likelihood with wearable technology. *Frontiers in neurology*, page 1170. 26, 48, 50, 52, 56, 57, 58, 60, 61

[Stirling et al., 2021c] Stirling, R. E., Maturana, M. I., Karoly, P. J., Nurse, E. S., McCutcheon, K., Grayden, D. B., Ringo, S. G., Heasman, J. M., Hoare, R. J., Lai, A., et al. (2021c). Seizure forecasting using a novel sub-scalp ultra-long term eeg monitoring system. *Frontiers in Neurology*, page 1445. 125

[Štrumbelj and Kononenko, 2014] Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665. 42

[Sun and Morrell, 2014] Sun, F. T. and Morrell, M. J. (2014). The rns system: responsive cortical stimulation for the treatment of refractory partial epilepsy. *Expert review of medical devices*, 11(6):563–572. 2, 23, 30, 67, 76, 77, 78, 126, 165, 175, 183, 192

[Sun et al., 2018] Sun, M., Wang, F., Min, T., Zang, T., and Wang, Y. (2018). Prediction for high risk clinical symptoms of epilepsy based on deep learning algorithm. *IEEE access*, 6:77596–77605. 60

[Sundararajan and Najmi, 2020] Sundararajan, M. and Najmi, A. (2020). The many shapley values for model explanation. In *International Conference on Machine Learning*, pages 9269–9278. PMLR. 165, 182

[Szegedy et al., 2013] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*. 165, 182

[Tang et al., 2021] Tang, S., Dunnmon, J., Saab, K. K., Zhang, X., Huang, Q., Dubost, F., Rubin, D., and Lee-Messer, C. (2021). Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *International Conference on Learning Representations*. 63, 64

[Taran et al., 2021] Taran, S., Adhikari, N. K., and Fan, E. (2021). Falsifiability in medicine: What clinicians can learn from karl popper. *Intensive Care Medicine*, 47(9):1054–1056. 121

[Tasker, 1998] Tasker, R. C. (1998). Emergency treatment of acute seizures and status epilepticus. *Archives of disease in childhood*, 79(1):78–83. 78, 166, 183

[Taussig et al., 2015] Taussig, D., Montavont, A., and Isnard, J. (2015). Invasive eeg explorations. *Neurophysiologie Clinique/Clinical Neurophysiology*, 45(1):113–119. 17

[Taylor and Rommelfanger, 2022] Taylor, L. and Rommelfanger, K. S. (2022). Mitigating white western individualistic bias and creating more inclusive neuroscience. *Nature Reviews Neuroscience*, pages 1–2. 126

[Teixeira et al., 2012] Teixeira, C., Direito, B., Bandarabadi, M., and Dourado, A. (2012). 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*, pages 6530–6533. IEEE. 34, 36, 59, 91, 109, 117, 163

[Teixeira et al., 2014a] Teixeira, C., Favaro, G., Direito, B., Bandarabadi, M., Feldwisch-Drentrup, H., Ihle, M., Alvarado, C., Le Van Quyen, M., Schelter, B., Schulze-Bonhage, A., et al. (2014a). Brainatic: A system for real-time epileptic seizure prediction. In *Brain-Computer Interface Research*, pages 7–17. Springer. 79, 166

[Teixeira et al., 2014b] Teixeira, C. A., Direito, B., Bandarabadi, M., Le Van Quyen, M., Valderrama, M., Schelter, B., Schulze-Bonhage, A., Navarro, V., Sales, F., and Dourado, A. (2014b). Epileptic seizure predictors based on computational intelligence techniques: A comparative study with 278 patients. *Computer methods and programs in biomedicine*, 114(3):324–336. 34, 35, 36, 50, 51, 52, 53, 54, 56, 58, 59, 61, 87, 100, 104, 161, 162, 163

[Thomas et al., 2020] Thomas, A. H., Aminifar, A., and Atienza, D. (2020). Noise-resilient and interpretable epileptic seizure detection. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE. 63, 64

[Timonen et al., 2018] Timonen, V., Foley, G., and Conlon, C. (2018). Challenges when using grounded theory: A pragmatic introduction to doing gt research. *International Journal of Qualitative Methods*, 17(1):1609406918758086. 123

[Tripathy et al., 2020] Tripathy, R. K., Ghosh, S. K., Gajbhiye, P., and Acharya, U. R. (2020). Development of automated sleep stage classification system using multivariate projection-based fixed boundary empirical wavelet transform and entropy features extracted from multichannel eeg signals. *Entropy*, 22(10):1141. 38

[Troshani and Wickramasinghe, 2014] Troshani, I. and Wickramasinghe, N. (2014). Tackling complexity in e-health with actor-network theory. In *2014 47th Hawaii International Conference on System Sciences*, pages 2994–3003. IEEE. 70

[Truong et al., 2019] Truong, N. D., Kuhlmann, L., Bonyadi, M. R., Querlioz, D., Zhou, L., and Kavehei, O. (2019). Epileptic seizure forecasting with generative adversarial networks. *IEEE Access*, 7:143999–144009. 50, 52, 54, 56, 57, 58, 60

[Truong et al., 2018] Truong, N. D., Nguyen, A. D., Kuhlmann, L., Bonyadi, M. R., Yang, J., Ippolito, S., and Kavehei, O. (2018). Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural Networks*, 105:104–111. 48, 49, 50, 52, 56, 58, 60

[Tsiouris et al., 2018] Tsiouris, K. M., Pezoulas, V. C., Zervakis, M., Konitsiotis, S., Koutsouris, D. D., and Fotiadis, D. I. (2018). A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals. *Computers in biology and medicine*, 99:24–37. 48, 50, 51, 52, 54, 56, 57, 58, 60

[Tsymbal et al., 2008] Tsymbal, A., Pechenizkiy, M., Cunningham, P., and Puuronen, S. (2008). Dynamic integration of classifiers for handling concept drift. *Information fusion*, 9(1):56–68. 36

[Usman et al., 2020] Usman, S. M., Khalid, S., and Aslam, M. H. (2020). Epileptic seizures prediction using deep learning techniques. *Ieee Access*, 8:39998–40007. 242

[Usman et al., 2021a] Usman, S. M., Khalid, S., and Bashir, S. (2021a). A deep learning based ensemble learning method for epileptic seizure prediction. *Computers in Biology and Medicine*, 136:104710. 242

[Usman et al., 2021b] Usman, S. M., Khalid, S., and Bashir, Z. (2021b). Epileptic seizure prediction using scalp electroencephalogram signals. *Biocybernetics and Biomedical Engineering*, 41(1):211–220. 48, 50, 51, 52, 54, 55, 56, 57, 58, 60, 61

[Uyttenhove et al., 2020] Uyttenhove, T., Maes, A., Van Steenkiste, T., Deschrijver, D., and Dhaene, T. (2020). Interpretable epilepsy detection in routine, interictal eeg data using deep learning. In *Machine Learning for Health*, pages 355–366. PMLR. 63, 64

[Valente et al., 2021] Valente, F., Henriques, J., Paredes, S., Rocha, T., de Carvalho, P., and Morais, J. (2021). A new approach for interpretability and reliability in clinical risk prediction: Acute coronary syndrome scenario. *Artificial Intelligence in Medicine*, 117:102113. 122

[Van Mierlo et al., 2014] Van Mierlo, P., Papadopoulou, M., Carrette, E., Boon, P., Vandenberghe, S., Vonck, K., and Marinazzo, D. (2014). Functional brain connectivity from eeg in epilepsy: Seizure prediction and epileptogenic focus localization. *Progress in neurobiology*, 121:19–35. 9

[Varsavsky et al., 2011] Varsavsky, A., Mareels, I., and Cook, M. (2011). *Epileptic seizures and the EEG: measurement, models, detection and prediction.* Taylor & Francis. 13, 16, 17, 26

[Vasileiou et al., 2018] Vasileiou, K., Barnett, J., Thorpe, S., and Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC medical research methodology*, 18(1):1–18. 113

[Venkatesh and Anuradha, 2019] Venkatesh, B. and Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and Information Technologies*, 19(1):3–26. 241

[Verma et al., 2020] Verma, S., Dickerson, J., and Hines, K. (2020). Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*. 123

[Viana et al., 2022] Viana, P. F., Nasseri, M., Duun-Henriksen, J., Biondi, A., Winston, J., Martins, P., Nurse, E., Dümpelmann, M., Worrell, G., Schulze-Bonhage, A., et al. (2022). Seizure forecasting using minimally invasive, ultra-long-term subcutaneous eeg: Generalizable cross-patient models. *Epilepsia*. 48, 50, 52, 54, 56, 57, 58, 60

[Vilamala et al., 2017] Vilamala, A., Madsen, K. H., and Hansen, L. K. (2017). 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)*, pages 1–6. IEEE. 63, 64

[Vuk and Curk, 2006] Vuk, M. and Curk, T. (2006). Roc curve, lift chart and calibration plot. *Advances in methodology and Statistics*, 3(1):89–108. 111

[Wachter et al., 2017] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841. 42, 111, 123, 165, 182

[Wang et al., 2018] Wang, D., Ren, D., Li, K., Feng, Y., Ma, D., Yan, X., and Wang, G. (2018). Epileptic seizure detection in long-term eeg recordings by using wavelet-based directed transfer function. *IEEE Transactions on Biomedical Engineering*, 65(11):2591–2599. 14

[Wang et al., 2017] Wang, G., Deng, Z., and Choi, K.-S. (2017). Detection of epilepsy with electroencephalogram using rule-based classifiers. *Neurocomputing*, 228:283–290. 63, 64

[Wang and Chen, 2019] Wang, Y. and Chen, Z. (2019). An update for epilepsy research and antiepileptic drug development: Toward precise circuit therapy. *Pharmacology & therapeutics*, 201:77–93. 18, 19

[Weisdorf et al., 2019] Weisdorf, S., Duun-Henriksen, J., Kjeldsen, M. J., Poulsen, F. R., Gangstad, S. W., and Kjær, T. W. (2019). Ultra-long-term subcutaneous home monitoring of epilepsy—490 days of eeg from nine patients. *Epilepsia*, 60(11):2204–2214. 18, 49, 125

[Weisstein, 2004] Weisstein, E. W. (2004). Bonferroni correction. *https://mathworld. wolfram. com/*. 256

[Wennberg, 2011] Wennberg, R. (2011). "introduction to eeg for nonepileptologists working in seizure prediction and dynamics. *Epilepsy: The Intersection of Neurosciences, Biology, Mathematics, Engineering, and Physics*, pages 23–39. 16, 17

[Wickramasinghe et al., 2007] Wickramasinghe, N., Bali, R. K., and Tatnall, A. (2007). Using actor network theory to understand network centric healthcare operations. *International Journal of Electronic Healthcare*, 3(3):317–328. 70, 71

[Wilkinson, 2013] Wilkinson, M. (2013). Testing the null hypothesis: the forgotten legacy of karl popper? *Journal of sports sciences*, 31(9):919–920. 121

[Winterhalder et al., 2003] Winterhalder, M., Maiwald, T., Voss, H., Aschenbrenner-Scheibe, R., Timmer, J., and Schulze-Bonhage, A. (2003). The seizure prediction characteristic: a general framework to assess and compare seizure prediction methods. *Epilepsy & Behavior*, 4(3):318–325. xxi, xxv, 28, 29, 30, 31, 32, 37, 60, 79, 93, 100, 109, 165, 177, 178, 180, 197

[Wolf et al., 2012] Wolf, P., Lin, K., and Nikanorova, M. (2012). Non-pharmacological therapy of epilepsy. *Oxford Textbook of Epilepsy and Epileptic Seizures*. xxix, 24

[Xu et al., 2020] Xu, Y., Yang, J., Zhao, S., Wu, H., and Sawan, M. (2020). An end-to-end deep learning approach for epileptic seizure prediction. In *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 266–270. IEEE. 48, 49, 50, 52, 54, 56, 57, 58, 60, 61

[Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer. 42

[Zhang et al., 2022] Zhang, J., Chatzichristos, C., Vandecasteele, K., Swinnen, L., Broux, V., Cleeren, E., Van Paesschen, W., and De Vos, M. (2022). Automatic

annotation correction for wearable eeg based epileptic seizure detection. *Journal of Neural Engineering*, 19(1):016038. 17, 49

[Zhang et al., 2020] Zhang, X., Yao, L., Dong, M., Liu, Z., Zhang, Y., and Li, Y. (2020). Adversarial representation learning for robust patient-independent epileptic seizure detection. *IEEE journal of biomedical and health informatics*, 24(10):2852–2859. 63, 64

[Zhang et al., 2019] Zhang, Y., Guo, Y., Yang, P., Chen, W., and Lo, B. (2019). Epilepsy seizure prediction on eeg using common spatial pattern and convolutional neural network. *IEEE Journal of Biomedical and Health Informatics*, 24(2):465–474. 48, 50, 51, 52, 53, 54, 56, 58, 60, 61

Appendix A

Other Contributions

During this thesis, I have also worked as an invited assistant professor in the Informatics Engineering Department of the University of Coimbra. I have also studied the prediction of Multiple Sclerosis disease progression using Machine Learning (ML) models. This chapter shows some of these parallel contributions.

Invited teaching assistant

P1 Teaching assistant in the Informatics Systems course, namely the laboratory classes, which were taught to Biomedical Engineering and Physics Engineering students in the following academic years: 2019/2020, 2020/2021, and 2021/2022.

Articles in International Journals

J11 **Pinto, M. F.**, Oliveira, H., Batista, S., Cruz, L., Pinto, M., Correia, I., Martins, P., Teixeira, C., "Prediction of disease progression and outcomes in multiple sclerosis with machine learning", *Scientific Reports* 10, 21038, DOI: 10.1038/s41598-020-78212-6 (2020).

Master's degree theses co-advisoring

J6 Oliveira, H.. "Evaluation and Prediction of Multiple Sclerosis Disease Progression", *Master Thesis dissertation, Universidade de Coimbra* (2020).

J7 Sousa, M.. "On the Explainability of Multiple Sclerosis Disease Progression Models", *Master Thesis dissertation, Universidade de Coimbra* (2021).

J8 Baião, J.. "On the Short-Term Prediction of Multiple Sclerosis Disease Progression", *Master Thesis dissertation, Universidade de Coimbra* (to be finished in 2022).

Summer and Training Schools

S1 Invited Speaker: "On the deconstruction and limitations of machine learning approaches to predict Multiple Sclerosis progression in real-life scenarios", PMSMatTrain Research Summer, 25-27 May in Denmark (2021).

S2 Participation in COST Action CA15140 Training School: Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO), Coimbra (2019).

Science communication to the general public

G4 Pinto, M., "Poeta professor de educação física", Douda Correria (2021). A poetry book about science.

G5 The "Introduction to JAVA programming language" and "Introduction to Machine Learning" workshops were given for several student associations, namely: Biomedical Engineering National Student Association (ANEEB), Junior Enterprise for Science and Technology (JEST), Students' Association of the Physics Department of several universities, such as the University of Lisbon, the University of Minho, and the University of Coimbra. These were all given between 2020 and 2021.

G6 Participation in the summer "jeKnowledge Academy 2022" to talk to high school students about ethical issues in using artificial intelligence, as well as current research in epilepsy EEG seizure prediction and prediction in multiple sclerosis disease progression.

Appendix B

Features description

This chapter provides a brief overview of the most extracted features from the state-of-the-art, and also provides a more detailed description of the used features for the Multiobjective Evolutionary Algorithm (MOEA) (chapter 5) and the explainability (chapter 6) studies. It is worth noting that, in these studies, only linear univariate features were used.

Linear features are mathematical techniques that use the signal's phase/frequency and amplitude information and comply with the linearity property. When one extracts linear features, assumes the quasi-stationarity of the Electroencephalogram (EEG) signal within each time window from the data segmentation step in the signal pre-processing stage.

Linear univariate measures

Within time-domain features, the first four statistical moments (mean, variance, skewness, kurtosis) characterise the amplitude distribution of the EEG time series [Kuhlmann et al., 2018a, Direito et al., 2017, Rasekhi et al., 2015, Teixeira et al., 2014b]. These are simple, light to compute, and can also be applied to other characteristics besides the electrical amplitude, as in the case of spectral information [Direito et al., 2017, Assi et al., 2017]. The skewness is zero for symmetric amplitude distributions and non-zero for asymmetric distributions. The kurtosis measures the relative peakedness or flatness of an amplitude distribution [Mormann et al., 2007]. These statistical measures have shown significant changes during the preictal period compared to the interictal state [Rasekhi et al., 2013, Mormann et al., 2005]. A decrease in the variance and an increase in the kurtosis were observed in the preictal period when compared with interictal based data [Teixeira et al., 2014b, Aarabi et al., 2009].

Hjorth parameters concern activity, mobility, and complexity, which are measures of mean power, root-mean-squared frequency, and root-mean-square frequency spread, respectively. These detect an intensification of brain activity which leads to an increase of energy [Kuhlmann et al., 2018a, Direito et al., 2017, Rasekhi et al.,

2015, Teixeira et al., 2014b, Moghim and Corne, 2014]. Authors have reported a significant increase in the mobility and complexity of the EEG during the preictal period [Rasekhi et al., 2013, Mormann et al., 2005].

Auto-regressive models and decorrelation time [Kuhlmann et al., 2018a, Direito et al., 2017, Rasekhi et al., 2015, Teixeira et al., 2014b, Rasekhi et al., 2013] have been used to inspect neural synchronisation. Auto-regressive models are used to model the EEG, where authors have used either the modelling coefficient values as features or the modelling error as a result of a seizure-generation process [Chisci et al., 2010]. Decorrelation time uses the autocorrelation function to find repeating patterns or identify the fundamental frequency hidden by its harmonic frequencies. Decorrelation time concerns the first zero-crossing of the autocorrelation function and provides information concerning the typical time scale of the data variability. This function can also be interpreted to measure signal stochasticity as a zero decorrelation time value means that a given signal is purely stochastic (white noise). Authors found that a decrease in the decorrelation time may detect a preictal period [Rasekhi et al., 2013, Mormann et al., 2005].

For capturing shifts from low to high frequencies, authors use features in the frequency domain. Several authors have decomposed the EEG into frequency bands (delta, theta, alpha, beta, and gamma) and computed their relative power. These are the most used features [Kuhlmann et al., 2018a, Direito et al., 2017, Rasekhi et al., 2015, Bandarabadi et al., 2015b, Assi et al., 2015, Teixeira et al., 2014b, Moghim and Corne, 2014, Rasekhi et al., 2013] and can be calculated by firstly computing the Power Spectral Density (PSD) of the time series within a time window. It is essential to mention that the calculation of the PSD assumes the signal in each window is short enough to be considered quasi-stationarity and long enough to capture the brain's low-frequency activity. Authors have reported that brain activity increases or decreases at specific frequency bands before seizures, where there may be a transfer of power from the lower to the higher frequencies [Bandarabadi et al., 2015b, Teixeira et al., 2014b, Rasekhi et al., 2013, Mormann et al., 2005]. For example, [Mormann et al., 2005] showed a decrease in Delta band power, coupled with a relative power decrease in the other sub-bands. [Bandarabadi et al., 2015b] showed that relative combinations of sub-band spectral powers across channel pairs might be used for tracking gradual changes preceding seizures.

Wavelet transform is a time-frequency domain transform that can be an alternative to the Fast Fourier Transform (FFT) as it decomposes the signal in different resolution levels according to different frequency ranges [Kuhlmann et al., 2018a, Direito et al., 2017, Teixeira et al., 2014b, Rasekhi et al., 2013]. In other words, wavelets provide a time-variant decomposition adapted to the signal, capturing minor details and sudden changes by providing higher frequency resolution for lower frequencies and higher time resolution to higher frequencies. With the signal decomposed, it is possible to compute several measures using the wavelet coefficients, as in the case of

signal energy. By computing the energy of the signals originated by the decomposition, a measure of the energy in different frequency ranges can be achieved [Teixeira et al., 2014b, Rasekhi et al., 2013, Gadhoumi et al., 2013].

Usually, most EEG signal's power is within the frequency band from 0 Hz up to 40Hz [Mormann et al., 2007]. Spectral edge frequency (SEF) is the measure that indicates the frequency below x per cent of the overall signal power is contained and is frequently used in seizure prediction [Direito et al., 2017, Teixeira et al., 2014b, Rasekhi et al., 2013, Mormann et al., 2005]. As the existence of a power transfer from low to high frequencies has been reported during the preictal stage, SEF may also be capable of capturing these dynamics.

Nonlinear univariate measures

The correlation dimension, the largest Lyapunov exponent, and the dynamic similarity index are frequent features derived from dynamic systems theory [Moghim and Corne, 2014, Mormann et al., 2005]. Chaotic measures can help explain brain dynamics as the EEG is a noisy and nonstationarity time series. Theoretically, a reduction of chaos may indicate impending seizures, as the predictability of brain dynamics tends to increase before a seizure. All these measures tend to capture the brain synchrony increase occurring before seizures. Additionally, seizures may be unexplainable using concepts of linear dynamics as these sometimes may be caused by external inputs [Iasemidis, 2003]. Nevertheless, nonlinear features may be too computationally expensive to be applied online [Assi et al., 2017].

The correlation dimension of a signal [Grassberger and Procaccia, 1983, Grassberger, 1983] measures the space dimension occupied by signal samples and is one of the various methods for fractal dimension assessment. Mathematically speaking, it provides an estimation of the complexity of attractors. The correlation sum quantifies the probability that two vectors of the state space trajectory lie within a given distance from each other.

According to the chaos theory, a system predictability is sensitive to the initial state conditions. The exponential divergence or convergence of nearby trajectories in the state space reflects the chaos inherent to a system [Moghim and Corne, 2014, Iasemidis, 2003]. The Lyapunov exponent [Kuhlmann et al., 2018a, Moghim and Corne, 2014, Mormann et al., 2005] is a measure of the system's chaotic behaviour as it quantifies the exponential divergence of two state-space trajectories that start close to each other. It can be computed by fitting an exponential regression.

Entropy [Teixeira et al., 2012] measures the regularity and the unpredictability of fluctuations over the EEG data. As a synchronous brain state characterises a seizure, entropy has been reported to detect changes from the interictal state to the preictal one.

Dynamic Similarity Index [Rabbi et al., 2013, Le Van Quyen et al., 1999] mea-

sures the similarity between segments of the EEG. It quantifies the difference in dynamics between a reference interictal EEG segment and a sliding window. The reference window is known to be considerably far from any seizure, containing the typical characteristics of the interictal activity. From it, one extracts nonlinear characteristics that are computed using time-delay embedding. Then, to each running window, the exact nonlinear characteristics are extracted. When the difference in these windows surpasses a given threshold, it is assumed to detect the preictal period.

Linear bivariate and multivariate measures

Bivariate and multivariate features characterise interactions between different regions of the brain and, therefore, different electrodes. As the preictal stage is a spatio-temporal complex state and seizures are known to be electrical discharges due to brain synchronisation, these features can capture and quantify this state [Assi et al., 2017, Rabbi et al., 2013].

The most used features are the maximum linear cross-correlation and multivariate autoregressive models, which are bivariate measures. Maximum cross-correlation measures the linear synchronisation of two electrode channels where a normalised value is obtained [Mormann et al., 2003]. When close to one, channels present a similar profile with a possible time lag. Other measures are possible, such as ratios between different spectral band powers [Bandarabadi et al., 2015b, Bandarabadi et al., 2015a].

Independent Component Analysis (ICA) [Oja and Hyvarinen, 2000] is another strategy for a signal decomposition where it assumes that each measured signal is a linear combination of independent signals and decomposes multidimensional data into statistically independent components. The components can be used to extract features. ICA can also be used for denoising [Assi et al., 2017, Acharya et al., 2013].

Nonlinear bivariate and multivariate measures

Bivariate and multivariate nonlinear measures have also been employed. By inspecting information in several electrodes simultaneously, these aim to capture synchrony changes using similarity and mutual information measures. One of the most used is mean phase coherence [Kuhlmann et al., 2018a, Rabbi et al., 2013, Mormann et al., 2005] that aims to quantify phase synchronisation between two channels. Dynamic entrainment [Iasemidis et al., 2004] aims to quantify the nonlinear behaviour of two electrode signals, which requires the estimation of the largest Lyapunov exponent, over time, for each analysed channel.

Appendix C

Ecosystem paper route

Here is presented how all the social network literature, from chapter 4, was selected.

Table SC.1 shows the references derived from the initial literature. Table SC.2 shows the papers selected due to the social network discussion with the research team. Google Scholar was the only used search engine.

Table C.1: The selected papers from the initial literature selection. As displayed, some of the papers referenced in these studies were also selected.

Authors	Topic	How it was selected
[Mormann et al., 2007]	Critic vision on seizure prediction	Initial literature
[Freestone et al., 2017]	Critic vision on seizure prediction	Initial literature
[Kuhlmann et al., 2018b]	Critic vision on seizure prediction	Initial literature
[Molnar, 2019]	ML Interpretability	Initial literature
[Schulze-Bonhage et al., 2010]	Patients vision on devices	Initial literature
[Ramgopal et al., 2014]	Survey on devices	Initial literature
[Winterhalder et al., 2003]	Academic studies	Mormann et al. 2007
[Schelter et al., 2006]	Academic studies	Mormann et al. 2007
[Andrzejak et al., 2003]	Academic studies	Mormann et al. 2007
[Cook et al., 2013]	Neurovista Clinica Trial	Freestone et al. 2017
[Sun and Morrell, 2014]	The RNS System	Freestone et al. 2017
[Gadhoumi et al., 2016a]	A review of methods	Freestone et al. 2017
[Nurse et al., 2016]	Real-life processing chip	Freestone et al. 2017
[Karoly et al., 2017]	Concept drifts	Kuhlmann et al. 2018
[Baud et al., 2018]	Concept drifts	Kuhlmann et al. 2018
[Doshi-Velez and Kim, 2017]	Explainability evaluation	Molnar C. 2019
[Lombrozo, 2006]	Explainability/Trust	Molnar C. 2019
[Ribeiro et al., 2016]	Explainability	Molnar C. 2019
[Lage et al., 2019]	Explainability	Molnar C. 2019
[Friedman, 2001]	Explainability	Molnar C. 2019
[Goldstein et al., 2015]	Explainability	Molnar C. 2019
[Apley and Zhu, 2020]	Explainability	Molnar C. 2019
[Ribeiro et al., 2018]	Explainability	Molnar C. 2019
[Sundararajan and Najmi, 2020]	Explainability	Molnar C. 2019
[Lundberg and Lee, 2017]	Explainability	Molnar C. 2019
[Bäck et al., 2020]	Explainability	Molnar C. 2019
[Szegedy et al., 2013]	Explainability	Molnar C. 2019
[Kim et al., 2016]	Explainability	Molnar C. 2019
[Cook, 1977]	Explainability	Molnar C. 2019
[Goodman and Flaxman, 2017]	GDPR Article 22	Doshi-Velez et al. 2017
[Wachter et al., 2017]	Explainability	Doshi-Velez et al. 2017

Table C.2: The selected papers from the initial literature selection. As displayed, some of the papers referenced in these studies were also selected.

Authors	Topic	Discussion with	Search Procedure or Search String
[Engel, 2016]	Pre-Surgical Monitoring	Team members	Shown by team member
[Becker et al., 2020]	SeizeIT2 clinical trial	Team members	Shown by team member
[Lewis, 2019]	Patient/Legislation	Team members	Shown by team member
[Chu et al., 2016]	Patient/Legislation	Team members	Shown by team member
[Rathore and Radhakrishnan, 2015]	Pre-Surgical Monitoring	Medical team	”epilepsy presurgical evaluation flowchart”
[Ben-Menachem, 2002]	Prospective Applications	Medical team	”vagus-nerve stimulation epilepsy”
[Boon et al., 2007]	Prospective Applications	Medical team	”deep brain stimulation epilepsy”
Foundation Epilepsy [Foundation, 2020]	Prospective Applications	Medical team	”deep brain stimulation FDA approval”
[Scheffer et al., 2017]	Brain Dynamics	Medical team	”ILAE epilepsy classification”
[Jafarpour et al., 2019]	Brain Dynamics	Medical team	”epilepsy seizure clusters definition”
[Tasker, 1998]	Emergency medication	Medical team	”status epilepticus emergency treatment”
[Gaínza-Lein et al., 2017]	Emergency medication	Medical team	”rescue medication epilepsy”
[Scheepers et al., 2000]	Emergency medication	Medical team	”seizure rescue medication”
[Dreifuss et al., 1998]	Emergency medication	Medical team	”rectal diazepam gel epilepsy”
[Foundation, 2020]	Emergency medication	Medical team	Searched in Epilepsy.com
[Assi et al., 2017]	A review of methods	Data science team	Shown by team member
[Debener et al., 2015]	Signal acquisition	Data science team	Shown by team member
[Teixeira et al., 2014a]	Real-life application	Data science team	Developed by the team
[Sisterson et al., 2020]	Testing the RNS	Data science team	”evaluating the RNS system”
[Schirrmeister et al., 2017]	Explainability in Deep Learning	Data science team	Shown by team member
[Islam et al., 2020]	Signal preprocessing	Data science team	”EEG seizure prediction ambulatory preprocessing”
[Beckers et al., 2021]	Prospective Applications	Reviewers	”EU medical device regulation”
[Majety et al., 2021]	Prospective Applications	Reviewers	”EU medical device regulation”

Appendix D

Ecosystem social network iteration and refinement details

The ecosystem social network iterations and refinement details, from chapter 4, are presented here.

Figures D.1-D.5 concern major iterations of the social network construction. Figure D.5 concerns the complex network obtained before refinement and encapsulation. Please note that some actors' names and numbers have been modified during the refinement stage.

During the network discussion, it was decided to add more details to some parts, as the explanation case. Evaluation levels, explanation range, and explanation strategies were added, which were found in Interpretable Machine Learning book [Molnar, 2019] and in related articles [Doshi-Velez and Kim, 2017]. Technological requirements and commercialisation were also detailed. The following were included: i) hardware aspects, such as recharging, heating, placement and removal, maintenance, price, client support, and fast processing, that can be found in [Ramgopal et al., 2014], and ii) information regarding General Data Protection Regulation (GDPR) article 22 that can be found by analysing [Doshi-Velez and Kim, 2017] and [Goodman and Flaxman, 2017]. The GDPR is a good refinement case, which concerns an actor found during the inspection of related articles within the initial ones [Molnar, 2019] until reaching saturation. It was also decided to highlight possible seizure interventions found in several initial papers [Kuhlmann et al., 2018b, Freestone et al., 2017, Ramgopal et al., 2014]. For the case of seizure interventions that deliver anti-epileptic drugs, this work got input from the clinician authoring this study regarding rescue medication such as diazepam. He advised us to search for epilepsy seizure rescue medication and stressed the importance of epilepsy clinical heterogeneity, which was also considered. Clusters of seizures (4.5) did not appear in the iteration models as they were included only in the system requirements. This was a codification limitation of this work which was successfully corrected by discussing the network among all authors of this study.

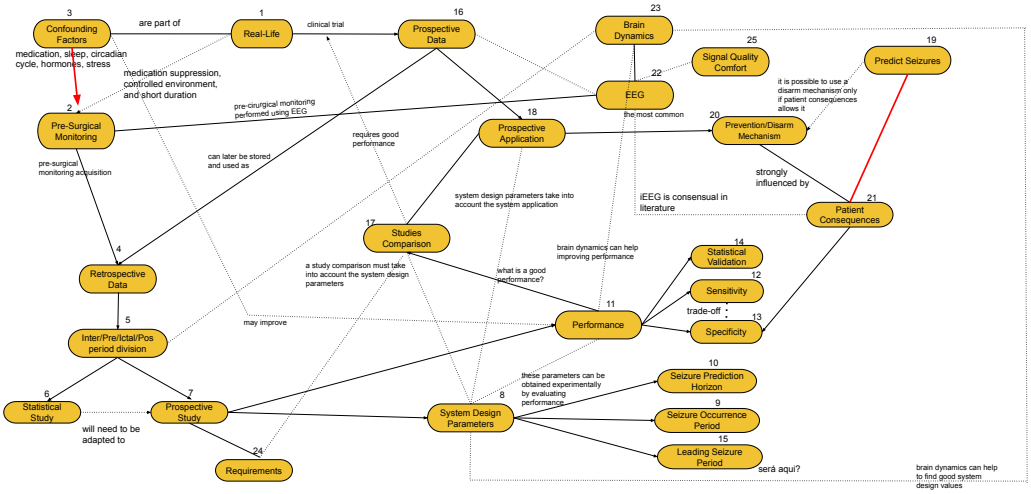


Figure D.1: Social network iteration after analysing [Mormann et al., 2007] and related articles. Red relations concern doubts raised at the time.

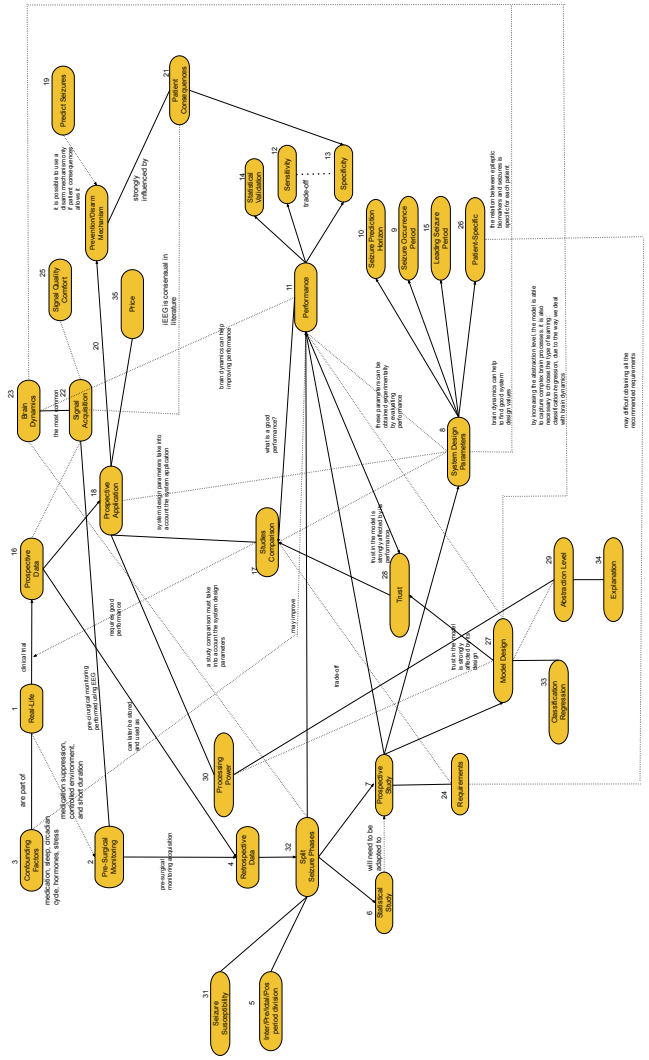


Figure D.2: Social network iteration after analysing [Freestone et al., 2017] and related articles.

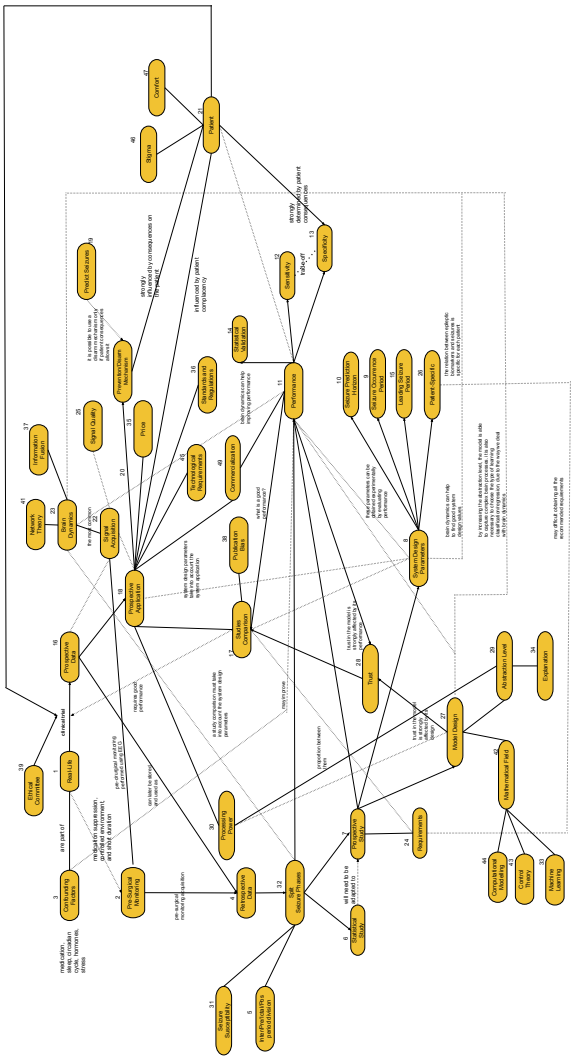


Figure D.3: Social network iteration after analysing [Kuhlmann et al., 2018b] and related articles.

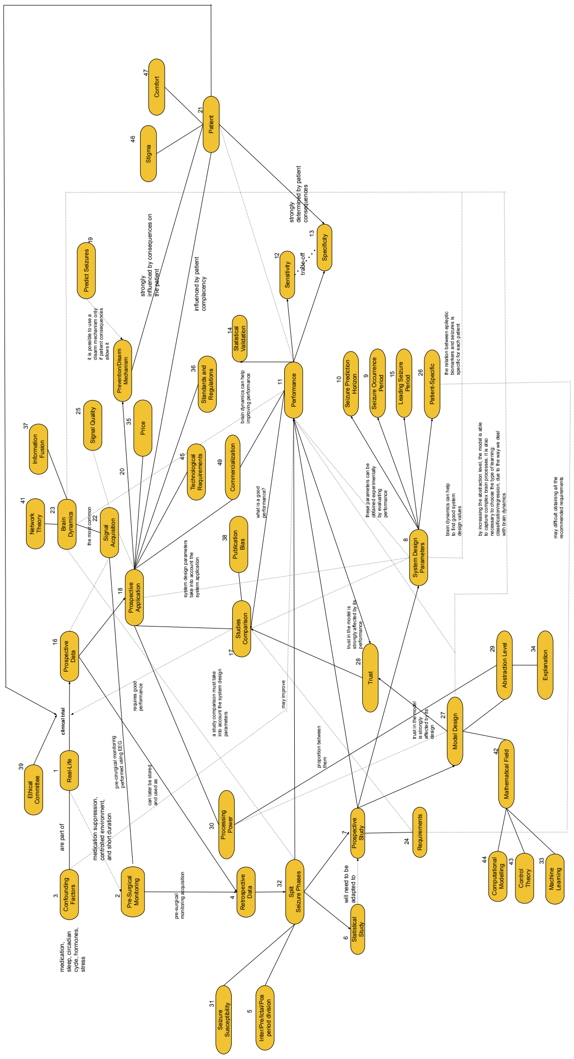


Figure D.4: Social network iteration after analysing [Ramgopal et al., 2014] and related articles.

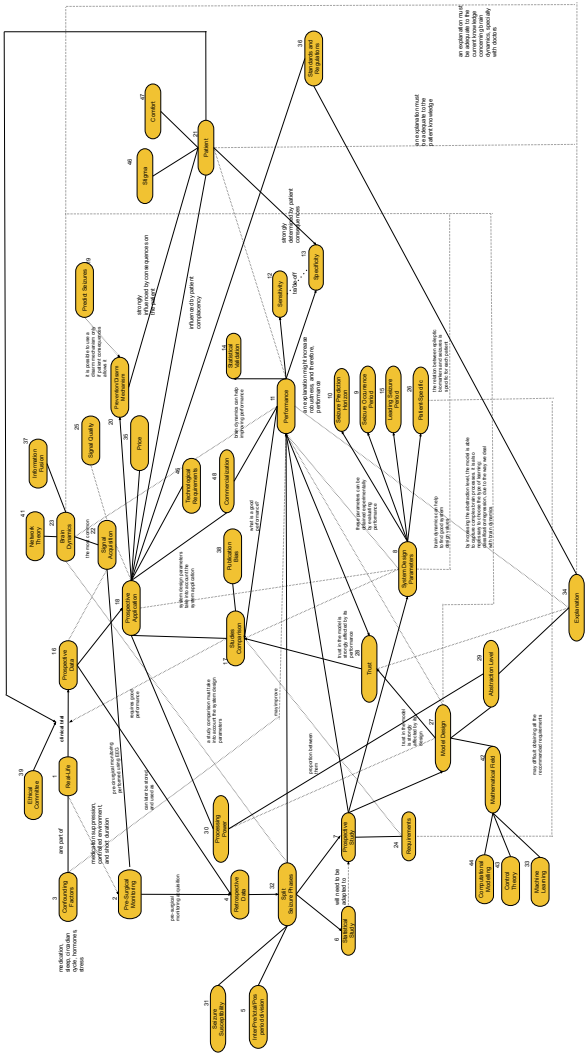


Figure D.5: Social network iteration after analysing Interpretable Machine Learning book [Molnar, 2019] and related articles. Technical aspects on explainability evaluation and range are not present simply due to the figure size.

Appendix E

The seizure prediction ecosystem

The seizure prediction ecosystem (the obtained network from chapter 4) is presented here in full detail, describing the relations between actors.

Actors (x) and relations (x-y) are named with numbers and grouped in colours to provide a better understanding. This section will explain these relations while deepening parts that require more detail. In the end, guidelines are provided to help authors design their research. An interactive version as Supplementary Material is also available, which allows a free exploration and may be more intuitive.

Real life and Presurgical Monitoring

This ecosystem begins with the real life of an epileptic patient (1). Years after being diagnosed with Drug-Resistant Epilepsy (DRE), a patient is referred to an epilepsy centre to undergo presurgical monitoring (5). The latter evaluates brain electrical activity (4) to localise the epileptic focus. If easily localised, removing the epileptic region is a possible solution [Engel, 2016, Mormann et al., 2007]. To perform this evaluation, one must perform signal acquisition (2), being the EEG the most commonly used signal (2–4). To acquire and study this data, patient consent (16→3) and an ethical justification (3) are required. In this case, there is a strong motivation. Please note that this is a simplification of the presurgical monitoring process. The following subsection provides a more detailed explanation.

Despite presurgical monitoring is not as frequent as desired, happening for less than 1% of DRE patients, most studies are performed using presurgical monitoring data. Therefore, this data may not represent real-life (2→5): the patient is in a controlled environment [Kuhlmann et al., 2018b, Freestone et al., 2017]; the patient body may take time to adapt to the acquisition material (as initial data may need to be discarded) [Cook et al., 2013]; clinicians suppress medication to increase seizure occurrence frequency; and the short period, typically, a couple of weeks of clinic

admission and signal recording [Baud et al., 2018, Alotaiby et al., 2014] may mask the influence (1- -5) of day-to-day confounding factors (6- -4), such as stress, circadian and ultradian rhythms.

Most databases comprise presurgical monitoring recordings, which correspond to retrospective data (7) that authors can indefinitely use in academic studies (8). To collect prospective data during a clinical trial in a real-life scenario (2→14), it is also necessary to find sufficiently strong and ethical motivation, which is discussed later. Briefly, prospective studies require a significantly higher patient complacency, involve longer periods, demand additional resources, and include higher risks for the patient. Prospective data then becomes retrospective (14- -7) [Kuhlmann et al., 2018b, Freestone et al., 2017].

Presurgical monitoring details

Presurgical monitoring aims to successfully localise, and delineate the extension of the epileptogenic zone, ideally followed by a surgery to remove it. Towards this, clinicians begin the patient analysis with a multimodal approach: long-term Electroencephalogram (EEG) and video recording, structural MRI, and neuropsychological evaluation. With this information, patients undergo resective surgery if: i) different approaches present coherent findings, ii) there is a well-defined epileptic region, and iii) there is a reasonable risk-benefit ratio.

When this process fails to identify and/or delineate the epileptic region, other signals can be acquired, such as magnetic source imaging (MSI), functional MRI, SPECT, and PET. With these, clinicians verify if there is a chance of generating a testable hypothesis regarding the epileptogenic zone. In a positive case, the patient will undergo intracranial EEG acquisition, cortical stimulation, and mapping. If the epileptogenic zone can be localised and resected, the patient will undergo surgery. Otherwise, antiepileptic drugs, ketogenic diet, or neurostimulation are the possible current solutions [Rathore and Radhakrishnan, 2015].

In the literature, one can find different studies using data acquired during presurgical monitoring collected using both scalp EEG and [Assi et al., 2017, Mormann et al., 2007]. Thus, when comparing EEG seizure prediction among different types of EEG, it is relevant to understand and consider the situation that leads to the Invasive Electroencephalogram (iEEG) acquisition.

One must not forget that a patient is referred to a level 3 or 4 epilepsy centre to do presurgical monitoring only after being diagnosed as drug-resistant, which can take many years after diagnosis, often too late to prevent irreversible damage cause by seizures. In fact, in the USA, fewer than 1% of DRE patients are examined by a multidisciplinary epilepsy team [Engel, 2016].

Brain Dynamics

Brain dynamics (4) play a fundamental role in predicting seizures. Ictogenesis is known for leading to a hyperexcitability state that increases brain synchronisation (see Figure E.1). Thus, the EEG (4.1.1) is the most used signal. It can be acquired using scalp or iEEG, each one addressing different assumptions on brain dynamics and therefore being more compatible with specific applications [Ramgopal et al., 2014, Mormann et al., 2007].

Scalp EEG obtains electrical activity from all surface regions, which is more suitable for handling the network theory (4.2.1): the latter proposes that seizures may arise from abnormal activity that results from a large-scale functional network and spans across lobes and hemispheres [Mormann et al., 2007]. Still, scalp EEG requires significant patient complacency as they cause stigma and discomfort. One can also expect frequent signal artefacts and noise. Its intervention application could be a warning system to reduce seizure consequences, which may be the most affordable option and, therefore, the one that requires fewer resources [Ramgopal et al., 2014]. Although iEEG has a higher signal-to-noise ratio and can be used to develop closed-loop intervention systems, patients may suffer from haemorrhage, device movement or infection, among others [Sun and Morrell, 2014]. Authors commonly focus on

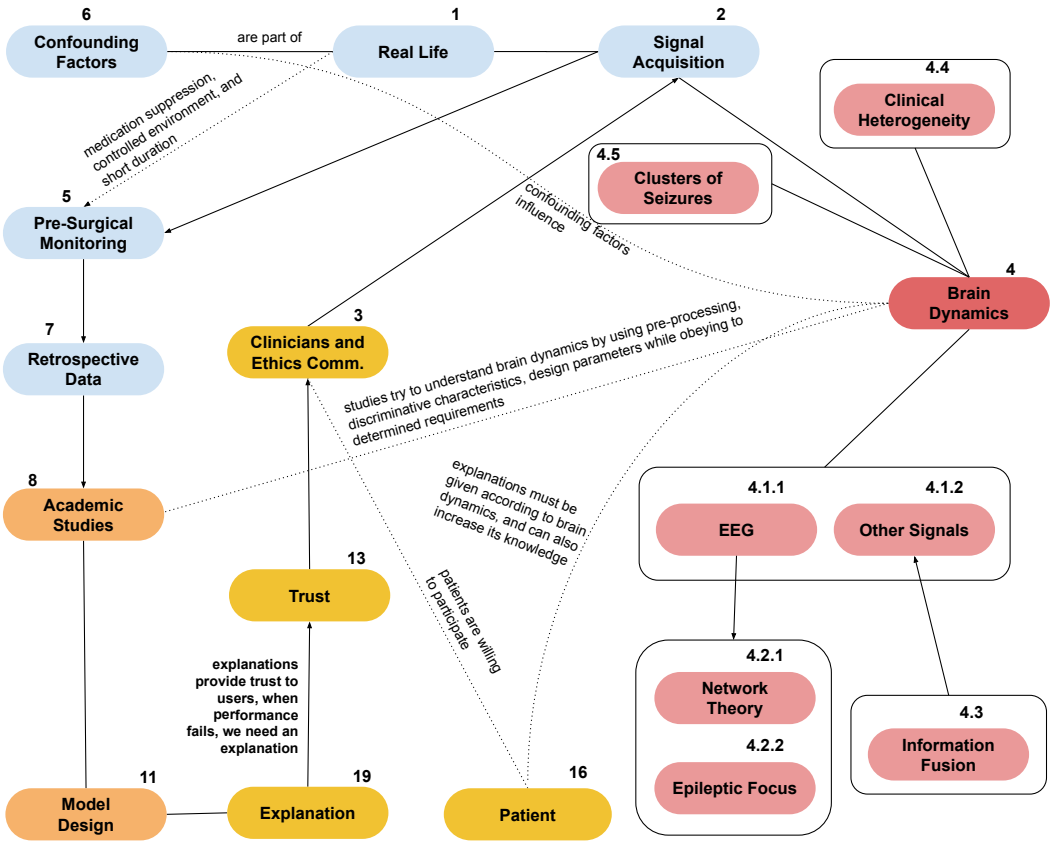


Figure E.1: Details on the relations between actors concerning brain dynamics. Non-major actors are inside boxes.

brain activity belonging to a given region, generally the epileptic focus (4.2.2). In fact, authors assume it is possible to predict seizures by only inspecting the epileptogenic area. Furthermore, the SeizeIT2 clinical trial [Becker et al., 2020] also explores EEG behind the ear that brings higher patient comfort, and [Debener et al., 2015] developed an EEG-ear array which demonstrated feasibility for long-term recordings. Other recent approaches, such as Minder from Epiminder and UNEEG, also capture EEG dynamics in given regions. However, these are invasive, and thus, side effects may be more severe.

Other sources of information (4.1.2) can be used to explore changes in brain dynamics (e.g., MRI) and also alterations in other non-neurological physiological parameters occurring during preictal interval [Kuhlmann et al., 2018b]. For example, the cardiovascular dynamics regulated by the autonomous nervous system can be captured by the electrocardiogram, which has been proven to carry complementary information for seizure prediction. Hence the growing belief that the analysis of multimodal data may provide improved results [Kuhlmann et al., 2018b]. In fact, multiple confirmations that the same dynamics may be present at different scales and biosignals (4.3) might enhance explainability and, therefore, increase trust (19→13), as mentioned in the following sections.

Moreover, the large clinical heterogeneity associated with epilepsy (4.4) also promotes current research to deepen understanding of this disease. Different types of epilepsy characterise several types of epilepsy syndromes. Clinicians distinguish epilepsy types according to the types of seizures, clinical history, EEG data and imaging features. Furthermore, several co-morbidities may arise, such as intellectual and psychiatric dysfunction [Scheffer et al., 2017]. Seizure generating mechanisms are specific for each patient, and each type of seizure [Kuhlmann et al., 2018b, Freestone et al., 2017, Assi et al., 2017], even though the source of spiking activity, for example, remains unclear [Mormann et al., 2007]. Additionally, it has been suggested that brain hyperexcitability induces a time dependency on seizures that leads to the occurrence of clusters of seizures (4.5) [Kuhlmann et al., 2018b,Freestone et al., 2017]. This aspect turns the ictogenesis process more complex and challenging to understand [Jafarpour et al., 2019].

Academic Studies

Academic studies attempt to discover relevant brain dynamics by, under some requirements, finding optimal signal processing strategies, predictive characteristics (further referred to as features), and accurate models (8- -4). The majority uses retrospective data because of its availability. In such cases, findings should be interpreted as a proof-of-concept to demonstrate that some methodologies may be more suitable, even though they still need to be tested in a real context [Kuhlmann et al., 2018b]. It is advised to inspect Figure E.2 to obtain more details on academic stud-

ies. Inevitably, researchers make several assumptions (see "Assumptions" section in this document for more information) when designing a new study. These may result from the used mathematical models, available data and other limitations, or even reflect the researcher's knowledge concerning brain dynamics (8- -4).

Authors attempt to predict seizures by assuming the existence of the preictal period. The latter is the transition between the normal brain state (interictal period) and a seizure (ictal period). It is possible to define the preictal period in two different ways (8.1). One approach assumes it as a point of no return (8.1.1), leading necessarily to a seizure [Mormann et al., 2007]. Another method is to envision it as a period of brain susceptibility (8.1.2) where a hyperexcitable state may not lead to a seizure [Freestone et al., 2017, Cook et al., 2013]. These hypotheses influence the experimental design significantly, as it may be more difficult to have a ground truth or, in other words, correct labelling on brain hyperexcitability when no seizure occurred. Thus, despite limiting the understanding of brain dynamics, the point of no return is commonly used in academic studies.

Studies have requirements (9), which constitute established assumptions among peers on data representativity of either real life or a trustful proof-of-concept. By fulfilling these requirements, authors assume the best possible simulation of a real

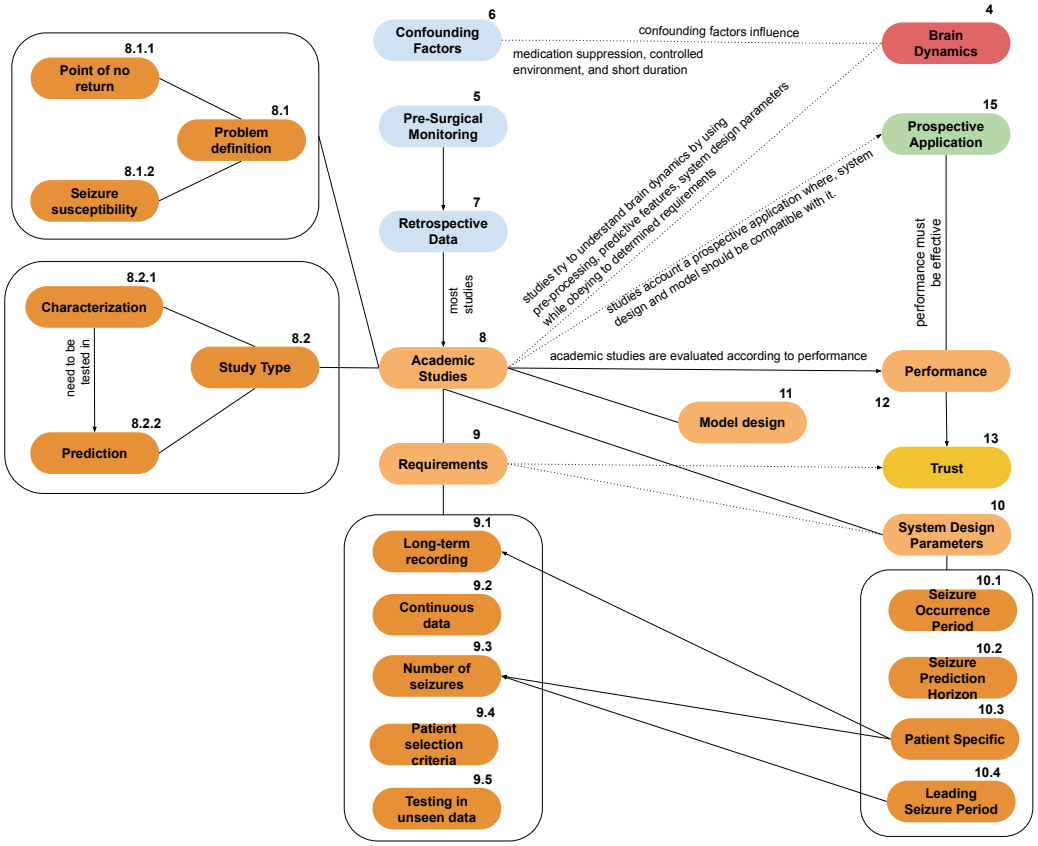


Figure E.2: Details on the relations between actors concerning academic studies. System design parameters are also commonly named as "seizure prediction characteristics" in the literature [Winterhalder et al., 2003]. Non-major actors are inside boxes.

context. The testing data requirements are: long term recordings (9.1), continuous data without manually removing any segments due to noise or artefacts (9.2), a minimum number of seizures to allow for training and testing of the models (9.3), rigorous patient selection criteria (9.4) where no patient was discarded based on performance, and models tested in unseen data (9.5) [Kuhlmann et al., 2018b, Gadhoumi et al., 2016a, Mormann et al., 2007].

It is relevant to note the existence of two types of studies (8.2): characterisation (8.2.1) and prediction (8.2.2) [Mormann et al., 2007]. In the first, authors try to find predictive models and/or features that capture a distinct behaviour between a normal brain state and the preictal period. However, the prediction potential should be further evaluated by integrating this information into a seizure prediction methodology (8.2.1→8.2.2) and observing the obtained performance. Prediction studies are the ones that simulate a real-life scenario and are designed to deliver timely interventions (8- -15). Therefore, these are the most reported in the literature and are the ones this work focus on here.

When considering a seizure intervention, system design parameters (10) have a significant role [Mormann et al., 2007, Winterhalder et al., 2003]. An alarm must be interpreted considering a Seizure Occurrence Period (SOP, 10.1), where a seizure is expected to occur, and a Seizure Prediction Horizon (SPH, 10.2), that guarantees time for an intervention. Furthermore, methodologies have converged for patient-specific algorithms (10.3) as authors have proven the existence of individual epileptic biomarkers. This influences study requirements (9- -10), as patient-specific strategies require a higher minimum recording duration (10.3→9.1) and a higher minimum number of seizures per patient (10.3→9.3). Finally, authors also must state the used seizure independence concept [Freestone et al., 2017] or, in other words, the minimum period necessary to assume that seizures have no relation (10.4). Due to brain excitability, consecutive seizures may occur in a short period. These create a cluster where the first seizure is the leading (and independent) one. It influences the number of independent seizures per patient (10.4→9.3) and limits the amount of data used. Note that there is no definition/rule to consider a seizure independent, representing another difficulty regarding brain dynamics (4). Additionally, it is worth noting that authors in prediction studies with presurgical monitoring data tend to use shorter periods [Assi et al., 2017] for defining seizure independence compared to a real-life scenario [Jafarpour et al., 2019].

Model Design

Figure E.3 shows detail concerning the design of mathematical prediction models. Seizure prediction entails the analysis of time series, which is typically initiated by segmenting into sliding windows. Thus, a seizure prediction model (11) might be able to distinguish brain states (interictal or preictal) throughout time. This model

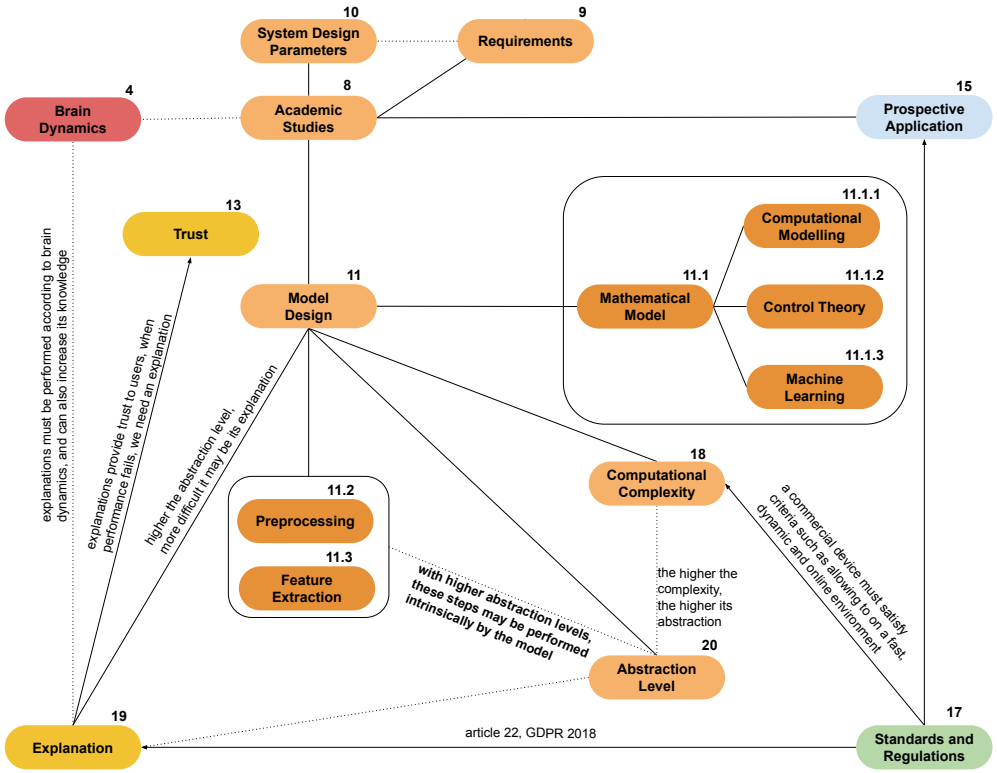


Figure E.3: Details on the relations between actors concerning model design. Non-major actors are inside boxes.

is a mathematical approach (11.1), which uses strategies from different domains, such as computational modelling (11.1.1), control theory (11.1.2), and the most common, machine learning (11.1.3), among others [Kuhlmann et al., 2018b, Assi et al., 2017, Gadhoumi et al., 2016a, Mormann et al., 2007].

Before training a model, authors may preprocess (11.2) the signals to remove noise while maintaining the frequencies of interest, and then they extract predictive features (11.3) [Assi et al., 2017, Gadhoumi et al., 2016a]. These two steps may be optional as more complex mathematical models have the theoretical potential to handle raw signals. A model, especially a machine learning one, can be distinguished by its abstraction level (20). Briefly, higher abstraction methods may intrinsically perform signal preprocessing (20- -11.2) and feature extraction (20- -11.3). Another relevant factor is computational complexity (18), where higher abstraction levels usually require higher processing power for algorithm development (18- -20). This can be an arising problem for real applications (17→18), as low computational requirements may be necessary [Kuhlmann et al., 2018b, Freestone et al., 2017].

Although not mentioned directly, by choosing a given preprocessing method, feature, and model, a researcher may be undertaking several assumptions on a physiological signal. Therefore, when constructing a pipeline, this work challenges authors to inspect them. Here is a list of questions one can ask: inside the chosen window length, can the, is the signal considered stationary, does it have noise, is it the result

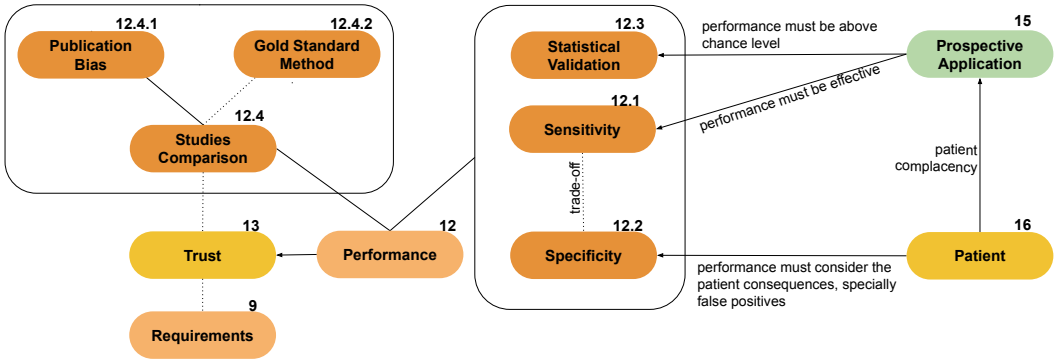


Figure E.4: Details on the relations between actors concerning performance. Non-major actors are inside boxes.

of linear interactions? Are the assumed brain dynamics simple or complex? Do they involve interactions? Although these may not change the experimental design, they can improve discussion and consequent comparison.

Performance

Performance is one of the most discussed aspects in seizure prediction studies (see Figure E.4). A promising methodology is naturally associated with model performance, which increases trust in the correspondent study (12→13). Sensitivity (12.1) corresponds to the ratio of correctly predicted seizures. Specificity (12.2) quantifies the number of false positives and is commonly obtained by counting the number of false alarms per hour [Gadhoumi et al., 2016a, Mormann et al., 2007, Winterhalder et al., 2003]. Statistical validation (12.3) [Kuhlmann et al., 2018b, Schelter et al., 2006, Andrzejak et al., 2003] has the goal of understanding if performance is above chance level as there is a trade-off between sensitivity and specificity (12.1- -12.2). In other words, this validation makes it possible to understand if the model's performance results from identifying random phenomena in the biosignals rather than seizure-related patterns. This aspect becomes more relevant considering that seizure prediction is a rare-event problem with a considerable imbalance between interictal and preictal intervals.

Some researchers suggest presenting an overall performance by computing the area under the receiver operating characteristic curve (relating false positive rate and true positive rate) [Kuhlmann et al., 2018b, Winterhalder et al., 2003]. However, the results can be interpreted according to the envisioned clinical application, specifically by considering patient intervention consequences (16→12.2). For instance, when considering the use of a warning system during presurgical monitoring, a maximum value of 0.15 FPR/h [Mormann et al., 2007, Winterhalder et al., 2003] has been considered as the upper limit of false alarms that cause bearable/tolerable levels of stress and anxiety.

Studies comparison (12.4) enables researchers to find acceptable methodologies

in different datasets and contexts while handling publication bias (12.4.1). This may occur when using retrospective data while trying several methodologies. When authors only report the best results and do not interpret failures as advances, their studies show overestimated performances or, in other words, overfitting to data [Kuhlmann et al., 2018b, Assi et al., 2017].

A proper comparison of studies requires more than comparing similar metrics. Authors are strongly recommended to use statistical validation to prove that the developed models overcome a random predictor in terms of performance [Mormann et al., 2007]. Nevertheless, it would be appropriate to compare results with a gold standard methodology applied in the same conditions [Kuhlmann et al., 2018b].

Trust and Explainability

After a proper studies comparison, one can ask what a good performance is or even inquire about the minimum performance that justifies the design of a clinical trial. A proper methodology is one which researchers trust. In literature, trust seems to be represented by studies reporting high performance (12→13) and complying with consensual study requirements (9- -13). By analysing data from longer recordings and/or a higher number of patients, trust increases as the testing data are more likely to represent real-life conditions [Kuhlmann et al., 2018b].

Although a given methodology eventually makes incorrect decisions, authors can still trust it if they can explain its decisions (19→13). A great scepticism concerning machine learning and high-level abstraction models may be due to the difficulty in delivering explanations about models' decisions [Molnar, 2019]. Although authors and/or clinicians are more willing to trust black-box models when they make correct decisions, wrong ones lead to mistrust because there is no human-comprehensible explanation [Freestone et al., 2017].

Trust should be a matter of concern when one designs a study. High-level abstraction models may have the potential to handle complex dynamics but require strong efforts toward providing explanations (19- -20). Current clinical knowledge of physiology should be the source of explanations as well as the basis for new findings (19- -4). As an explanation is an exchange of beliefs [Lombrozo, 2006], its acceptance may differ among patients, clinicians, and data scientists. To better understand trust and explainability, there is the need to inspect Figure E.5.

Explainability evaluation (19.1) is required. It is possible to evaluate an explanation on three levels: application (19.1.1), where it must satisfy an expert (e.g. a clinician and a data scientist); human (19.1.2), where it must explain the decision to a person with no field knowledge (e.g. a patient); and proxy (19.1.3) by establishing concrete criteria (e.g. the depth of a decision tree). The proxy level is the one requiring fewer resources. Nevertheless, it should be used with great care when a model has not proved its quality in delivering explanations, both at the application

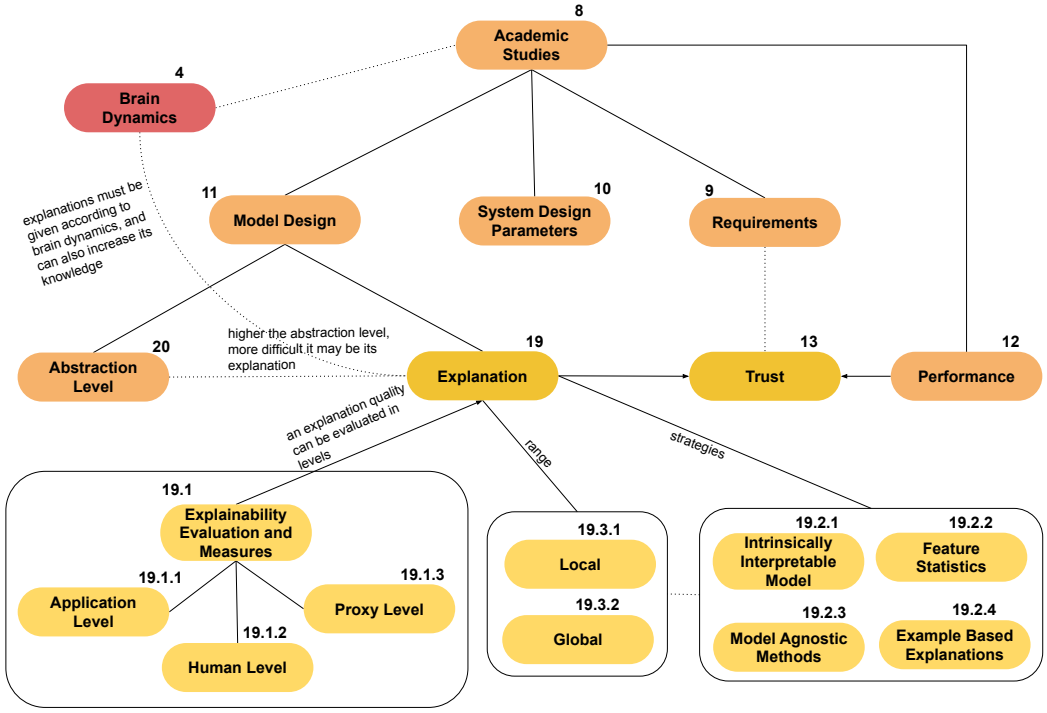


Figure E.5: Details on the relations between actors concerning trust and explainability. Non-major actors are inside boxes.

and human levels [Molnar, 2019, Doshi-Velez and Kim, 2017].

There are several strategies [Molnar, 2019, Lage et al., 2019] to retrieve an explanation which can be grouped in: i) intrinsically interpretable models (19.2.1) with a reduced set of features (such as decision trees, generalised linear models, k-NN, among others); ii) feature statistics (19.2.2) summary and visualisation; iii) agnostic methods (19.2.3), which work on top of developed models [Sundararajan and Najmi, 2020, Apley and Zhu, 2020, Ribeiro et al., 2018, Lundberg and Lee, 2017, Ribeiro et al., 2016, Goldstein et al., 2015]; and iv) example-based (19.2.4) by representing determined samples and showing the model decision [Bäck et al., 2020, Wachter et al., 2017, Kim et al., 2016, Szegedy et al., 2013]. The explanation range is also a topic of concern. It is local (19.3.1) when it only explains a given decision for a sample and respective neighbourhood [Molnar, 2019]. If it explains all samples, it is global (19.3.2).

Note that a possible relation between patient and trust (16→13) was not considered, as it concerns the algorithm design solely. Additionally, any connection between patient and explanation (16→19) was not mentioned directly, despite considering that a patient has the right to an adequate explanation concerning the device decisions. In fact, such rights are covered on article 22 from 2018 General Data Protection Regulation (GDPR) [Goodman and Flaxman, 2017, Doshi-Velez and Kim, 2017, Wachter et al., 2017]. In this case, an explanation and trust concern field experts, such as data scientists and clinicians. Nevertheless, patient comfort, trust

and a proper explanation are fundamental. Therefore, this work implicitly included these on the relation from the patient to the ethics committee (16→3), represented by the act of volunteering. When a patient volunteers, he/she demonstrates trust in researchers and clinicians, having already shown commitment to his/her well-being and ensured an adequate explanation.

Prospective Data and Applications

A methodology can be clinically approved (3→2 and 2→14) after years of research when it becomes trustworthy to experts and patients are willing to volunteer. Studies are trustworthy when they report high performance and good explainability while fulfilling all data requirements. It is possible to inspect details concerning prospective applications with Figure E.6.

Ideally, studies using retrospective data envision and open the way to the enrollment in potential prospective scenarios (8- -15) [Ramgopal et al., 2014, Mormann et al., 2007]. It is also possible to undergo a clinical trial without any seizure intervention, as it happens with the ongoing SeizeIT2 clinical trial (NCT04284072) and the ongoing Epiminder clinical trial (ACTRN12619001587190) to develop Minder. These studies may not achieve the goal of disarming a seizure yet, but they provide valuable data for authors, which may be seen as a good compromise between patient safety and research progress. Furthermore, from a non-prediction perspective, these studies can also improve the standard of care for people with epilepsy.

A prospective application has an intervention mechanism (15.1), which could be integrated in a closed-loop system, as is the case of vagus nerve stimulation (15.1.1) [Ben-Menachem, 2002], responsive cortical stimulation as with the RNS®system (15.1.2) [Sun and Morrell, 2014], or deep brain stimulation (15.1.3) [Boon et al., 2007]. The last was recently approved by the FDA [Foundation, 2018] and encompasses two ongoing trials (NCT03900468, NCT02076698). An alternative could be a warning system (15.1.4) designed to minimise seizure consequences [Ramgopal et al., 2014, Mormann et al., 2007] and/or taking seizure rescue medication, as benzodiazepines (15.1.5) [Gaínza-Lein et al., 2017, Scheepers et al., 2000, Tasker, 1998]. Selecting an adequate intervention strategy is complex and must account for patient complacency and consequences (16→15).

It is interesting to reflect on the ideal scenario [Mormann et al., 2007]. The development of a constant and effective intervention (15.2), such as chronic or scheduled stimulation from implantable devices, without any side effects (stress and anxiety, prolonged exposure to medication) and device-related problems (infection, intracranial haemorrhage, tissue reaction, skin erosion, lead migration, among others) would change the paradigm. Academic prediction studies would focus on increasing knowledge of brain dynamics (15.2→8) as there was no need to investigate another prospective application. Given the quantity of today's limitations, this may

be utopic. However, it may be relevant to stress the purpose of seizure prediction research.

Naturally, device manufacturers must obey industry standards and regulations (17→15) related to hardware safety aspects (15.2), such as recharging and low-energy consumption (15.2.1), heating (15.2.2), placement and removal (15.2.3), and maintenance (15.2.4). Other factors, equally important, concern an affordable price (15.3) and permanent client support (15.4). Consequently, the design of the models should consider the use of fast processing methods allowing its integration in small devices (17.1) [Kuhlmann et al., 2018b, Ramgopal et al., 2014]. It is essential to mention that considerable advances have been made in these devices, which is the case of IBM's neuromorphic TrueNorth chip [Nurse et al., 2016] that already allows for the deployment of deep learning models.

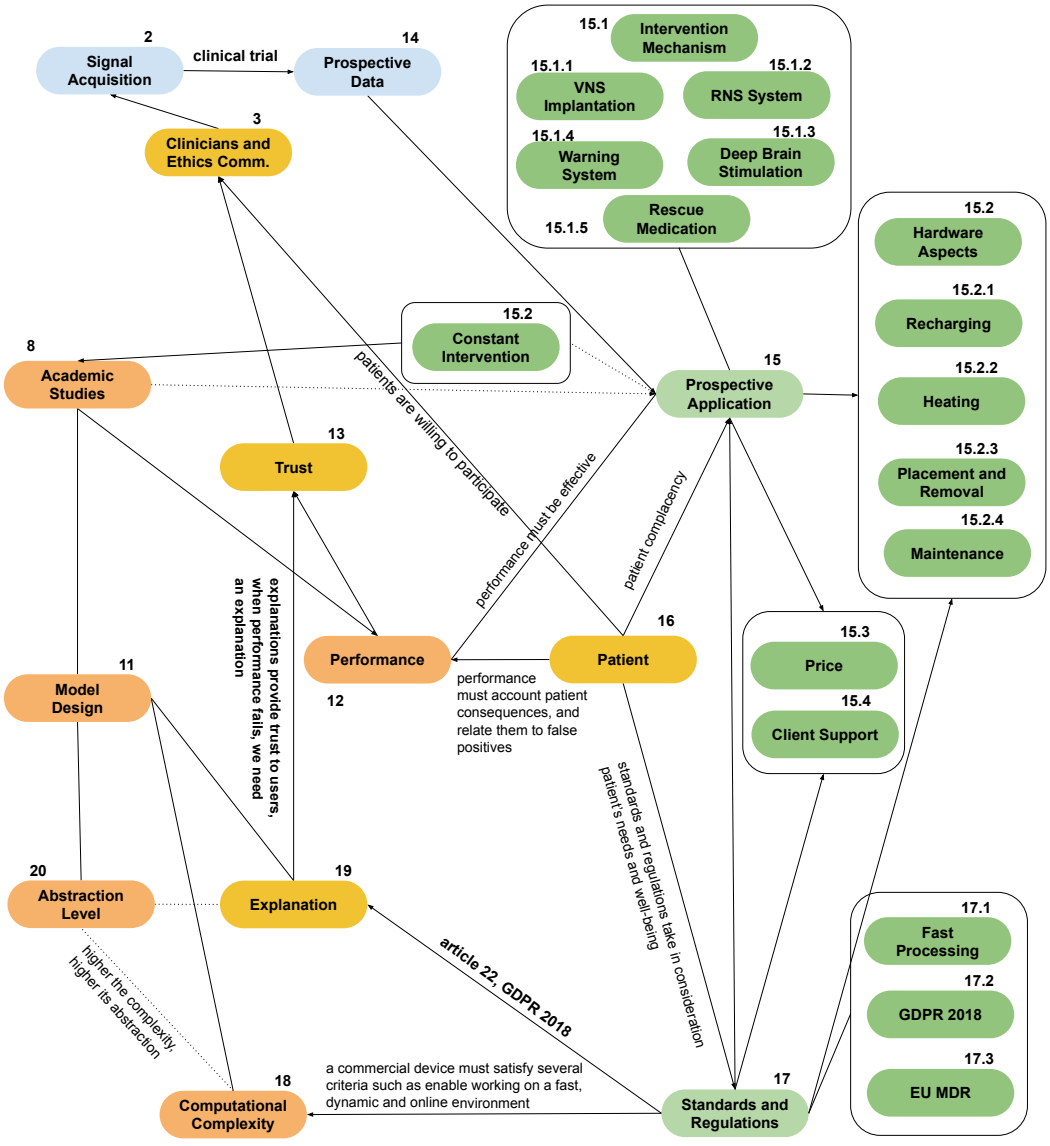


Figure E.6: Detail on the relations between actors concerning a prospective application. Non-major actors are inside boxes.

The price may be fundamental to the industry. Electrostimulation by implanting iEEG electrodes is currently considered the most promising strategy, as both Responsive Neurostimulation System (RNS) system and Neurovista's system used iEEG [Kuhlmann et al., 2018b, Freestone et al., 2017, Cook et al., 2013, Mormann et al., 2007]. However, these may demand higher human and monetary resources than presurgical scalp EEG monitoring, which is already inaccessible to a large part of DRE patients. In the United States of America (USA), for example, fewer than 1% of DRE patients are examined by a multidisciplinary epilepsy team. Besides, several only have access to level 3 or 4 epilepsy centres many years after onset, often too late to prevent irreversible damage caused by seizures [Engel, 2016]. Thus, by focusing immediate efforts on low-cost and accessible warning systems followed by rescue medication intake, it may be possible to reach considerably more DRE patients.

The 2018 GDPR [Goodman and Flaxman, 2017] (17.2) and the 2021 European Union Medical Device Regulation (EU MDR)) [Beckers et al., 2021] (17.3), for European citizens and European economic space, are also an important aspect. Article 22 presents the first steps towards legislation on algorithm explainability for high-risk decisions based on personal data (17→19). Thus, standards and regulations orientate authors toward patient safety (16→17).

Appendix F

Assumptions on a prediction study

The most common assumptions used in the seizure prediction field, concerning the ecosystem study from chapter 4, are presented here.

Assumptions are a crucial part of any study in any scientific field. Researchers often need to make assumptions about the world. The authors may use a different perspective and, therefore, different assumptions depending on the study question. In established areas, such as the seizure prediction field, authors may consider several assumptions that are not stated directly or not addressed properly in the discussion section. These assumptions may subconsciously be considered part of the public domain, particularly among peers. For non-experienced researchers, this may be a critical aspect.

Although there is the need to make assumptions to solve a problem, these should be periodically reviewed. Table SF.1 presents the major assumptions often adopted by authors. These concern the used data, signal acquisition, problem definition, types of studies, requirements, system parameters, and model design. Note that this list might not be complete as other topics can be missing, e.g., assuming a postictal period (a brain refractory period) or defining a period of adaptation of the brain to the seizure prediction device hardware.

Finally, an author must pay attention to all the assumptions made to verify if there are inconsistencies. For example, with an Invasive Electroencephalogram (iEEG), electrostimulation is usually the envisioned intervention. Thus, as the Responsive Neurostimulation System (RNS) system performs discharges up to 5000 ms, Seizure Occurrence Period (SOP) periods must be short. If an author uses scalp EEG instead, a warning system is the envisioned intervention. Thus, Seizure Prediction Horizon (SPH) periods must be significant to allow an intervention or medication intake followed by time to take effect.

There are assumptions concerning the used mathematical tools that must be accounted for, as well. These can be related to preprocessing, feature extraction,

and/or model training. For example, when using only a deep convolutional neural network, authors assume that the algorithm can automatically train a robust model while learning discriminative features and dealing with noise.

Another example regarding feature selection: by using filtering methods (such as the absolute value of Pearson correlation), researchers assume that features have independent discriminative power and, therefore, choose the features with the highest discriminative power. The best feature is chosen with a regularisation method (such as the LASSO regression) instead of the individual best. With a regularisation method, authors also account for the interaction between features by choosing the group with the highest discriminative power (these may not have a high individual discriminative power). Thus, biologically speaking, regularisation methods assume the possibility of existing more complex interactions in the brain when compared to filtering methods.

Table F.1: Major assumptions on seizure prediction studies. Others are also possible, especially the ones concerning mathematical operations.

Approach	Assumption
Data	
Using presurgical monitoring data	Presurgical monitoring data is either representative of real-life or constitutes a good proof-of-concept of it.
Signal Acquisition	
Scalp EEG	Seizure generation mechanisms may occur in any place of the brain. Supports the network theory. A warning device is envisioned.
Intracranial iEEG	Seizure generation mechanisms can be detected by inspecting only a given region, usually the epileptogenic focus. An invasive application, such as electrostimulation, is envisioned.
Other EEG method	The used method is more suited for a real application (as patient comfort) while ensuring effective performance.
Other physiological signals	The used method is more suited to a real application (providing more comfort to the patient) while capturing non-neurological seizure related dynamics. For example, the ECG signal.
Problem definition	
Preictal period	There is a point of no return in the brain after a seizure will always occur.
Using seizure susceptibility	There is a brain susceptibility period where hyperexcitability and synchronisation are probable. It may not lead to a seizure.
Fixed preictal period	All seizures are generated in an equal window of time.
Study types	
Characterisation	A good performance represents a proof-of-concept for potential use in a prediction study.
Prediction	A good performance constitutes a proof-of-concept for potential use in a clinical application.
Study Requirements	
Long-term continuous recordings, and testing in unseen data	These conditions represent a good proof-of-concept of a real application scenario.
Number of Seizures	The number of seizures is enough to represent real-life or to constitute a good proof-of-concept.
System Parameters	
Patient-specific models	Seizure generation mechanisms vary among patients.
Not using patient-specific models	Seizure generation mechanisms are similar among patients.
Specific models for each stage of circadian and or multidian rhythms	Circadian and or multidian rhythms influence seizure generation mechanisms.
Using the same model for all stages of circadian and or multidian rhythms	Circadian and or multidian rhythms do not influence seizure generation mechanisms.
Specific models for specific epilepsy syndromes, epilepsy types, medication, and so forth	The selected factors influence seizure generation mechanisms.
Using the same model for all epilepsy syndromes, epilepsy types, medication, and so forth	The selected factors do not influence seizure generation mechanisms.
SOP and SPH	Seizure generation mechanisms occur necessarily within the period determined from SOP+SPH to SPH, before seizure onset.
SOP	The used seizure occurrence period has an adequate duration to make an intervention effective.
SPH	The used seizure prediction horizon allows time enough to render the envisioned intervention possible.
Model Design	
Preprocessing	The acquired signals have artefacts and noise that can be removed with preprocessing.
Feature Extraction	It is possible to extract more robust measures of signal dynamics that characterize a pre-seizure state.
Mathematical model training	It is possible to develop a mathematical model that discriminates a normal brain state and a pre-seizure one.

Appendix G

Questions about the seizure prediction future

The questions made about the future of seizure prediction, concerning the ecosystem study from chapter 4, are presented here.

Explanations and trust

Explanations help detect data bias while increasing the robustness of the seizure prediction models. They are important to improve patient safety. They also help to mitigate scepticism regarding machine learning methodologies. Based on this, the following questions appeared, which must be handled by data scientists and clinicians:

1. Which are the concerns on explainability when designing seizure prediction models for prospective testing? Are clinicians sceptic about how the models work? Or are they afraid to compromise patients' safety? Do clinicians and data scientists have different needs concerning human-comprehensible explanations, or are these equal?

2. When compromising patient safety is the only main problem with non-human interpretable systems, do data scientists need to work on delivering deep explanations of the ictogenesis process? Or can they opt to improve some other parts of their methodology, e.g., increasing model robustness against data bias and noise?

Explanations and clinical approval

The need for explanations may justify that all clinically approved studies, such as the phase IV Neuropace Responsive Neurostimulation System (RNS) system (NCT00572195) and the phase I NeuroVista Seizure Advisory System (NCT01043406), use algorithms with features that are clinically intuitive [Freestone et al., 2017].

These two clinical trials demonstrate that, despite all the literature efforts put into developing complex models and consequent increase in performance, it may be necessary for a fully explainable model to provide trust. Secondly, the Seizure Advisory System clinical trial demonstrates the possibility of using models that are not necessarily intrinsically interpretable, as long as they produce human-comprehensible explanations while ensuring patient safety, handling data bias detection, and dealing with model robustness. Based on this, the following questions arose, which must be handled between data scientists and clinicians:

4. If those new approaches have satisfactory performance on the application and human levels, can they be used?

5. Do researchers need a human-comprehensible explanation when the algorithm is being used in real-time? Or do they need it only at certain moments, as with raised alarms and incorrect decisions? This may handle the fact that data scientists tend to trust model decisions when they are correct and only tend to inspect errors. In fact, when training a machine learning model, the training algorithm minimises misclassified samples' errors.

6. Can counterfactual explanations be interesting? Counterfactual explanations are very human-friendly and used widely by humans in daily life because they can answer a "why" question. This question can be formulated [Molnar, 2019] as: what is the slightest change to the features that would change the prediction from alarm to no-alarm?

7. The used features in these studies [Sun and Morrell, 2014, Cook et al., 2013] (line-length, bandpass, and energy-related measures) are clinically intuitive, and many others have been widely used in the literature, such as decorrelation time, Hjorth parameters, relative spectral power, wavelet decomposition, auto-correlation measures, auto-regressive modelling coefficients, and entropy. Which ones could also be used in a clinical trial?

8. These studies have used clinically intuitive features as input in the decision models. With the proper guarantee of model robustness, data bias detection, and patient safety, could Deep Learning approaches, with raw data as input, be used in clinical trials to perform feature extraction automatically? In a positive scenario, would the authors' methods need to explain which features were extracted by the Deep Learning model, or could an explanation just show the relevant data points for a given decision?

Patients and real-applications

In a survey [Schulze-Bonhage et al., 2010] on Drug-Resistant Epilepsy (DRE) patients concerning seizure prediction devices, patients expressed their preference for

an invasive solution. Acceptable performance concerned high values, with an SOP of 10 minutes, which, by inspecting literature, is currently not achievable, to the best of this thesis' knowledge. This study was mostly a fixed questionnaire with few open questions on these parameters (Seizure Occurrence Period (SOP), Seizure Prediction Horizon (SPH), and minimum performance) and preferences. For example, what would be an acceptable SOP duration? The options were: 10 minutes, 30 minutes, 1 hour, 3 hours, or more than 3 hours. The possibility of biasing answers is significant, which must be stressed. These led us to several questions which must be handled by data scientists, clinicians, and patients:

8. Could researchers obtain a different patient point of view with the same subjects if they undergo a different approach, such as open questions only followed by a grounded theory analysis?

9. Despite their preferences, do patients have the financial resources to acquire a seizure intervention device? Can the study be biased towards people with significant money resources? Do patients know the success rate of such applications? Are they truly aware of all possible consequences and problems (infections, haemorrhage) with implantable invasive systems and their chance of happening? Moreover, the latter may lead to even higher monetary and psychological costs.

10. Concerning scalp Electroencephalogram (EEG), few patients are willing to use long term acquisition systems. Should researchers make efforts in other formats of EEG scalp acquisition, such as the two-electrode system from SeizeIT2 [Becker et al., 2020] or the ear-EEG array [Debener et al., 2015]? Or should they focus on other signals, despite having a lower theoretical potential, such as the electroencephalogram (ECG)? For instance, smartwatches are more comfortable and can record a one-channel ECG. Are these strong reasons to promote enrolling in long-term clinical trials using these devices instead? They certainly allow more comfort and mitigate stigma, but their prediction performance might be not as good.

11. Patients claim to accept, as minimum performance and SOP duration, values that are not achievable, at least yet, in literature [Kuhlmann et al., 2018b,Freestone et al., 2017, Mormann et al., 2007] (10 minutes of SOP, minimum sensitivity of 90%, and very low False Prediction Rate per hour (FPR/h), simultaneously [Schulze-Bonhage et al., 2010]). If they knew more about current research, could they change their minds? Regarding an invasive solution with electrostimulation, is it relevant to have a low false alarm rate if electrical stimuli may not represent great harm? Note that in this case, additional problems of device heating or energy consumption are being excluded.

12. Should authors investigate the maximum false alarm rate that a patient can hold without large physical and/or psychological consequences (due to too much electrostimulation or medication intake) concerning all intervention systems? Is there another alternative to evaluate specificity quality?

The only commercial intervention device: The RNS system

The RNS system reduces seizure frequency over time. Nevertheless, patients still suffer seizures. Thus, the following question appeared, addressed to clinicians and data scientists:

13. Why do patients continue to have seizures? When a patient suffers a seizure, are these devices acting too late, during points of no return, or are they not detecting any preictal activity at all? Efforts have already been made towards a proper system evaluation [Sisterson et al., 2020]. Would these electrostimulation systems benefit from using more robust algorithms to predict these sooner, or are there seizures that brain electrostimulation can not disarm?

Appendix H

Ecosystem's guidelines extrapolation to forecasting

The possible extrapolation of the ecosystem study guidelines (from chapter 4) to seizure forecasting is presented here.

Despite this study's emphasis on seizure prediction, these guidelines and conclusions can be adapted and, thus, hold for seizure forecasting. Here, there are some adaptions to the guidelines that one may need to perform when performing a seizure forecasting study.

- Guideline 1: deciding to perform prediction or forecasting is already undertaking determined assumptions on brain dynamics, which should be highlighted in every study.
- Guideline 2: in seizure forecasting, researchers also need to envision a determined intervention (warning device, neuromodulation, cortical stimulation, rescue medication) to understand if the obtained system can be applied to real-life and how it will affect the patient (physically and mentally).
- Guideline 3: shifting from seizure forecasting to seizure prediction, at the lights of machine learning and in practical terms, might be a change in labels or to a regression problem. Even in the theoretical absence of clinically developed methodologies for forecasting, one can adapt the existing prediction machine learning methodologies to the appropriate labels.
- Guideline 4: forecasting methodologies also need to focus on explainability to promote trust among experts.

Appendix I

An acceptable performance in a prediction study for a clinical setting

The discussion about the accepted performance for a clinical setting, concerning chapter 4, is presented here.

An accepted performance for the clinical setting might depend on the chosen application. In terms of sensitivity, the minimum justifiable level might be subjective. However, in the paper that seeks the Drug-Resistant Epilepsy (DRE) Patients' views on seizure prediction devices [Schulze-Bonhage et al., 2010], patients claim a minimum of 90% performance for sensitivity. So that might be a reasonable limit to account for in warning devices.

However, researchers must not forget the limitations of defining this value for other applications, such as neurostimulation. Due to how the way the Responsive Neurostimulation System (RNS) system works, it is impossible to measure sensitivity performance. Nevertheless, although it may provoke too many stimulation interventions, it significantly reduces the seizure rate in many patients. Thus, it is clinically accepted.

The False Prediction Rate per hour (FPR/h) also depends on the chosen application. This value must be adapted to the envisioned intervention due to patient complacency.

For warning devices, there is a maximum established FPR/h value of 0.15 [Winterhalder et al., 2003] in pre-surgical monitoring that was calculated using their mean seizure rate in those conditions. Researchers can adapt that FPR/h value to real life by using the mean seizure frequency in those conditions.

When a neurostimulation intervention is envisioned, one must study the maximum intervention rate that a patient can hold without causing significant damage. When envisioning a rescue medication, such as benzodiazepines, one needs to understand the drugs' pharmacokinetics, their long-term side effects, and the maximum

frequency and dose intake.

Appendix J

MOEA study's patient data

The metadata about all analysed patients in chapter 5 is presented here.

Table J.1: Patient information. For each patient, it is presented its EPILEPSIAE ID, the number of seizures used for training and testing and their recording duration, seizure focus (as lobe-subegion-side, frontal lobe (f), temporal lobe (t), central lobe (c), occipital lobe (o), parietal lobe (p), mesial subregion (m), basal subregion (b), polar subregion (p), right side (r), left side (l), and bilateral (b)), the annotated activity pattern (unclear (?), rhythmic sharp waves (s), rhythmic alpha waves (a), rhythmic delta waves (d), rhythmic theta waves (t), rhythmic beta waves (b), repetitive spiking (r), cessation of inter-ictal activity (c), amplitude depression (m), low amplitude fast activity (l)), seizure classification (unclassified (UC), Focal Onset Aware (FOA), Focal Onset Impaired Awareness (FOIA), Focal to Bilateral Tonic-Clonic (FBTC)), sleep stage at onset (awake (A), sleep stage I (1), II (2), III (3), IV (4) and REM (R)), and surgery decision (performed (s), offered but not performed (o), not offered (n), and invasive monitoring required (i)).

Patient ID Hospital	**Sex Age**	**Fs**	**#Seiz. (Train/ Test)**	**Rec. duration (hours)**	**Focus loc.**	**Surgery decision**	**Seizure activity pattern**	**Seizure classification**	**Vigilance state at seizure onset**
102	m	256	3	12.00	c-l, p-r	i-s	t, t, t /	FOIA, FOIA, UC /	A, 2, A /
UKLFR	36		8	109.36			d, t, t, t, t, d, t, t	FOIA, UC, FOIA, FBTC, FOIA FOIA FOIA, FOIA	A, A, A, A, 2, A, 2, A
202	m	256	3	12.00	–l	s	b, r, r	FOIA, FOIA, FOIA	R, A, 2
UKLFR	24		1	22.12			t	FOIA	2
500	m	512	3	12.00	t-r	s	a, a, a	FOIA, FOIA, FOIA	A, A, A
HUC	57		4	66.39			a, a, a, a	FOIA, FOIA, FOIA, UC	A, A, A, A
1200	f	1024	3	12.00	t-r	o	t, a, a	FOIA, UC, FOIA	A, A, A
HUC	40		1	11.44			a	UC	A
1500	m	1024	3	12.00	t-r	o	t, t, t	FOIA, FOIA, FOIA	3, A,3
HUC	50		1	6.91			t	FOIA	A
1600	m	512	3	12.00	t-l	o	a, a, a	FOIA, UC, FOIA	2, 3, A
HUC	33		1	10.87			a	FOIA	A
2100	m	512	3	12.00	t-l	o	t, t, t	UC, FOIA, FOIA	2, A, A
HUC	20		3	36.16			t, t, t	FOIA, FOIA, FOIA	3, A, A

2300	m	512	3	12.00	t-l	o	t, t, t	FOA, UC, UC	A, 1, A
HUC	49		1	7.13			t	FOA	A
3300	m	1024	3	12.00	-	o	t, d, d	UC, UC, FOA	A, 2, A
HUC	29		1	8.50			t	UC	A
4500	m	1024	3	12.00	t-l	o	t, a, a	FOIA, FOIA, FOIA	A, A, A
HUC	49		1	14.62			a	FOIA	A
5800	f	1024	3	12.00	t-b	o	a, a, a	UC, UC, FOIA	A, A, A
HUC	26		2	18.35			a, a	UC, FOIA	A, A
6000	m	512	3	12.00	t-l	o	s, a, a	UC, UC, FBTC	A, 2, 3
HUC	16		1	48.71			d	UC	2
7200	m	1024	3	12.00	t-b	-	b, b, d	FBTC, FBTC, FBTC	2, 2, 2
HUC	41		1	4.17			d	FBTC, FBTC	2
8100	m	1024	3	12.00	t-b	n	b, b, b	FOIA, FOIA, UC	A, A, A
HUC	16		2	45.29			b, s	FOIA, UC	A, 1
8902	f	256	3	12.00	tpl	n	a, b, a	UC, FOIA, FOIA	A, A, A
UKLFR	67		2	22.48			m, a	FOIA, FOIA	A, A
11002	m	256	3	12.00	tpr	s	?, s, a	UC, FOIA, FOIA	A, R, A
UKLFR	41		1	1.69			t	FOIA	A
11502	m	256	3	12.00	tmr	s	b, b, b / b, a, ?, t, s, b, b, b, b, b	FOIA, FOA, FOIA / FOIA, FOA, FOA, FOA, UC, UC, FOA, FOIA, FOIA, FOIA	2, A, A / A, A, ?, A, 1, A, A, A, A, A
UKLFR	34		10	120.40					
12702	f	256	3	12.00	tbl, –b,	o	r, r, r	FOIA, UC, FOIA	A, 2, A
UKLFR	16		3	67.50	tll		l, l, l	UC, FOA, FOIA	2, A, 2

21602 UKLFR	m 44	256	3 6	12.00 57.75	tpr, f-l	o	?, ?, b / t, b, s, b, d, b	FOA, UC, UC UC, UC, UC, UC, UC, UC	A, A, A/ 2, A, 2, A, R, A
21902 UKLFR	m 47	256	3 1	12.00 9.76	t-l	s	t, t, t b	UC, FOIA, FOIA FOIA	A, A, A R
22602 UKLFR	m 26	256	3 2	12.00 22.03	f-l	i-n	b, b, ? b, b	UC, UC, UC UC, UC	A, A, A 2, A
23902 UKLFR	m 36	256	3 2	12.00 33.92	t-l	s	t, t, t d, t	FOA, FOA, FOA FOA, FOA	A, A, A A, A
26102 UKLFR	m 65	256	3 1	12.00 22.56	t-l	i-n	m, t, t t	FOIA, FOIA, FOIA FOIA	A, A, A A
30802 UKLFR	m 28	256	3 5	12.00 61.67	t-r, t-l	o	t, t, t t, t, t, t, t	FOA, FOA, FOA FOA, FOA, FOA, FOA, FOA	R, A, 2 A, A, R, 2, 2
32502 UKLFR	m 46	256	3 3	12.00 29.01	t-b, f-r	i-n	t, t, t t, t, t	FOIA, FBTC, UC UC, FBTC, FBTC	A, A, A A, 2, A
46702 UKLFR	f 15	256	3 2	12.00 12.88	tmr	o	a, a, t b, t	FOA, FOIA, FOIA FBTC, FOIA	A, 2, A 2, A
50802 UKLFR	m 43	256	3 2	12.00 35.58	t-l	i-n	t, t, t t, t	FOIA, UC, UC FOIA, FBTC	A, 2, 2 2, A
51002 UKLFR	f 46	256	3 5	12.00 47.10	f-r, t-l	s	?, a, a t, t, t, t, t	FOA, FOIA, FOIA FOIA, FOIA, FOIA, UC, FOIA	A, A, A A, A, A, A, A
52302 UKLFR	f 61	256	3 1	12.00 6.84	t-l	i-s	?, ?, d t	UC, FOA, UC UC	A, A, 1 A

53402	m	256	3	12.00	t-l, t-r	s	?, ?, ?	FOA, FOA, FOA	A, A, 2
UKLFR	39		2	50.45			?, t	FOA, FOIA	A, A
55202	f	256	3	12.00	t-r, t-b	i-n	t, d, t	FOIA, FOIA, FOA	A, A, A
UKLFR	17		5	65.32			t, t, t, r, r	UC, UC, FOA, UC, FOIA	A, A, A, A, A
56402	m	256	3	12.00	t-r, t-l	n	t, ?, ?	UC, UC, UC	A, A, A
UKLFR	47		1	20.23			a	FBTC	A, A
58602	m	256	3	12.00	t-l	i-n	r, t, t	FOIA, FOIA, FOIA	A, R, A
UKLFR	32		3	23.33			r, r, t	FOIA, FOIA, FOIA	A, A, 2
63502	f	1024	3	12.00	tml, t-r	o	s, s, s	FOIA, UC, FOIA	A, A, A
UKLFR	63		2	28.25			s, s	FOIA, FOIA	2, 2
64702	m	256	3	12.00	tmr	i-s	?, m, t	FOA, FBTC, FBTC	A, A, A
UKLFR	51		2	31.60			t, t	FBTC, FBTC	A, 2
70302	m	2500	3	12.00	tpr	s	s, ?, a / s, b, s, s, s, s, s	FBTC, FOIA, UC / UC, FOA, UC, FBTC, FBTC, UC, FBTC	2, 2, A / A, A, 2, A, 2, 2, 2
UKLFR	18		7	88.85					
70902	f	2500	3	12.00	-	n	?, ?, t	FOA, FOA, FOA	A, A, A
UKLFR	28		2	17.87			?, t	FOA, FBTC	A, A
71102	f	2500	3	12.00	t-r, t-l	n	t, b, t	FBTC, FBTC, FOA	2, A, A
UKLFR	31		2	16.65			r, r	FOA, FOA	A, A
71502	f	2500	3	12.00	–l, t-r	n	t, t, d	FOIA, FOA, UC	2, 2, 2
UKLFR	18		2	31.18			t, ?	SP, FBTC	2, 2
71802	f	2500	3	12.00	-	n	t, t, t	FOIA, FOIA, UC	A, A, A
UKLFR	46		2	33.76			?, t	FOA, FOIA	A, A

73002	m	2500	3	12.00	t-r, t-l	n	d, ?, t	FBTC, FBTC, FBTC	A, A, A
UKLFR	32		2	14.77			t, t	FOIA, UC	R, A
75102	f	256	3	12.00	-	n	t, a, t	UC, UC, UC	2, 2, 2
UKLFR	2		2	26.97			?, t	UC, UC	A, 2
75202	m	256	3	12.00	c-r, o-r,	s	t, t, t	FOIA, FOIA, UC	2, 2, A
UKLFR	13		4	52.62	t-r		t, t, ?, t	FOIA, FOIA, FOIA, FOIA	A, A, A, A
79502	f	256	3	12.00	p-l, t-l	i-n	s, b, r	FOIA, FOIA, FOIA	A, A, A
UKLFR	35		4	41.63			b, p, b b	FOIA, FOIA, FOIA, FOIA	A, A, A, A
80602	f	256	3	12.00	f-l	i-n	?, ?, r	FOIA, FOIA, FBTC	A, A, A
UKLFR	34		2	88.14			?, ?	FOIA, FOIA	A , A
80702	f	256	3	12.00	t-b	n	b, b, ?	FOIA, FOIA, UC	A, A, 2
UKLFR	22		3	29.53			c, c, c	FOIA, FBTCB, FOIA	2, 2, 2
81402	f	256	3	12.00	f-l	n	t, ?, t	UC, FOA, UC	A, A, 2
UKLFR	65		3	60.59			t, t, d	FOIA, FOIA, FOIA	2, 2, A
85202	f	256	3	12.00	t-l	s	m, c, m	FOIA, FOIA, UC	2, A, A
UKLFR	54		2	20.41			m, m	UC, UC	A, A
92102	m	256	3	12.00	plr	s	r, r, ?	FOIA, UC, FOIA	A, A, A
UKLFR	38		2	84.74			r, r	FOIA, FOIA	A, A
93402	m	256	3	12.00	tpl	n	t, t, t	FBTC, FOIA, FOIA	2, 2, 2
UKLFR	67		2	54.03			t, t	UC, UC	2, 2
93902	m	256	3	12.00	t-r	i-n	t, t, d	FOA, FOIA, FBTC	A, A, 2
UKLFR	50		3	20.27			d, d, d	FOIA, FOIA, UC	A, 2, A

94402	f	256	3	12.00	t-r	o	?, d, b	FOA, UC, FOIA	A, A, A
UKLFR	37		4	30.65			t, ?, b, ?	UC, FOA, UC, FOA	2, A, 2, A
95202	f	256	3	12.00	t-l	s	b, b, b	FBTC, FOIA, FOIA	2, 2, 2
UKLFR	50		4	90.25			m, b, b, t	FOIA, UC, FOIA, UC	2, 2, 2, 2
96002	m	256	3	12.00	t-l, t-r	s	t, t, t	FOIA, FOIA, FOIA, FOIA	A, A, A
UKLFR	58		4	82.14			d, a, t, a	UC, FOIA, FOIA	A, A, A, A
98102	m	256	3	12.00	t-l	i-n	?, ?, ?	FOA, UC, UC	A, A, A
UKLFR	36		2	45.73			?, ?	UC, FBTC	A, A, A, A
100002	m	256	3	12.00	-	o	?, t, ? /	FOA, FOA, FOA /	A, A, A /
UKLFR	35		6	72.19			t, t, c, c, ?, c	FOA, FOIA, FOA, UC, FOA, UC	A, 2, A, 1, A, 2
101702	m	256	3	12.00	t-r, t-l	n	t, t, t	FOIA, FOIA, FOIA	A, A, A
UKLFR	52		2	23.81			r, r	FOIA, FOIA	2, A
102202	m	256	3	12.00	t-l, tpl	n	b, ?, t	FOA, UC, FOIA	2, A, 2
UKLFR	17		4	51.39			?, t, t, t	UC, FOA, FOIA, UC	A, A, 2, A
103002	f	256	3	12.00	o-r	o	t, d, ?	FOIA, FOIA, FOIA	A, A, A
UKLFR	31		3	92.49			a, r, b	UC, FOIA, UC	A, A, A
103802	f	256	3	12.00	-	-	r, ?, t	FOIA, FOA, FOIA	2, A, A
UKLFR	17		3	19.64			t, t, t	FOIA, FOIA, FOIA	1, A
104602	f	256	3	12.00	tml	o	t, a, t	FOIA, FBTC, FBTC	2, A, A
UKLFR	17		2	15.24			t, d	FBTC, UC	1, A
109202	f	256	3	12.00	f-l, f-r	n	?, a, a	FOIA, FBTC, FBTC	A, 2, R
UKLFR	49		1	7.20			b	FBTC	2

109502	m	256	3	12.00	t-l, t-r	n	t, t, t	FOIA, FOIA, UC	A, A, A
UKLFR	50		1	41.91			t	UC	A
109602	f	1024	3	12.00	tml	s	b, r, b	FOIA, FOIA, FOIA	A, A, A
UKLFR	32		1	21.29			b	UC	A
110602	m	256	3	12.00	t-r	o	t, t, t	FOIA, FOIA, FOIA	A, A, A
UKLFR	56		2	25.90			t, t	FOIA, FOA	A, A
111902	m	256	3	12.00	c-l	n	d, t, t	FBTC, UC, FBTC	A, A, A
UKLFR	42		2	14.57			s, t	FBTC, FBTC	A, A
112402	f	256	3	12.00	f-r	i-s	d, t, ? /	FOIA, FOA, FOIA /	A, A, A /
UKLFR	32		9	61.80			b, b, b, d, t,	FOIA, FOIA, FOIA, FBTC,	2, A, A, A,
							b, s, b, t	FOIA, FOIA, FOA, UC, FOIA	A, A, A, A 1
113902	f	256	3	12.00	t-r	o	t, d, t	UC, FOIA, FOIA	A, A, 2
UKLFR	29		3	22.73			t, t, t	FOIA, UC, FOIA	A, 2 A
114702	f	256	3	12.00	tpr	o	t, t, t	FOIA, FOIA, UC/	A, A, A/
UKLFR	22		5	34.02			t, d, t, d, t	FOIA, FOIA, FOIA, FOIA	A, A, A, A
								FOIA	A
114902	f	256	3	12.00	t-r, tll	n	s, b, s	FOA, FOIA, FOIA	A, A, A
UKLFR	16		4	50.63			t, r, a, t	FBTC, UC, FOIA, FOIA	2, A, A, A
115202	f	256	3	12.00	o-r, t-l	n	t, t, t	FOIA, UC, FOIA	2, 2, 2
UKLFR	32		4	70.96			t, b, b, b	FOIA, FOIA, FOIA, UC	2, 2, 2, 2
123902	f	256	3	12.00	t-l, t-r	i-n	t, t, t	FBTC, FBTC, FOIA	2, 2, R
UKLFR	25		2	30.15			t, t	FOIA FOA	A, A

125002	m	256	3	12.00	-	n	?, ?, ?	UC, FOIA, UC	A, A, A
UKLFR	39		4	39.28			?, ?, ?, a	FOIA, FOIA, UC, UC	A, 2, A, A
1235103	m	400	3	12.00	t-b	n	r, l, l	UC, UC, UC	3, A, ?
HdlPS	19		2	34.71			l, l	UC, FBTC	3, 3
1307103	f	512	3	12.00	t-b, t-l,	o	l, l, l	FOIA, FOIA, FOIA\	A, A, 1\
HdlPS	37		5	191.76	h-l		l, s, t, t, l	FOIA, FOIA, FOIA, FOIA,	A, A, A, A
								FOIA	A
1308503	m	512	3	12.00	h-l, t-l	-	s, l, l	FOIA, FOIA, UC	2, A, A
HdlPS	19		1	8.98			l	FOIA	2
1308603	m	512	3	12.00	t-b, t-r	-	t, t, t	UC, FOIA, FOIA	1, A, A
HdlPS	35		1	64.49			m	FOIA	A
1310803	f	512	3	12.00	tml	o	t, t, t	FOIA, FOIA, FOIA	?, ?, A
HdlPS	38		1	22.28			t	FOIA	A
1312903	f	512	3	12.00	fll	i-n	a, t, a	FOIA, FOIA, UC / UC	2, A, 1
HdlPS	19		1	48.01			t		1
1315203	m	512	3	12.00	t-r, f-r	o	l, d, t	UC, FOIA, UC / FOIA	R, A, R
HdlPS	21		1	34.81			t, t		A
1319203	m	400	3	12.00	–r	n	l, l, m	UC, UC, UC	A, 1, A
HdlPS	28		4	65.70			m, a, a, a	FOIA, FOIA, UC, FOA	2, A, A, A
1321803	f	512	3	12.00	tmr, tmb,	i-n	t, t, d	UC, FOIA, FOIA	3, 2, A
HdlPS	27		1	4.17	tml		d	FOIA	2
1322803	m	400	3	12.00	t-r, c-r	i-n	t, t, t	FOIA, FOIA, FBTC	A, A, A
HdlPS	18		1	22.11			t	UC	A

1324903	f	400	3	12.00	c-r	o	l, l, l	UC, UC, FOA	1, 1, 1
HdlPS	22		2	17.43			l, l	UC, UC	A, A
1325103	f	400	3	12.00	t-r	o	a, l, l	FBTC, FOIA, UC	2, 2, 2
HdlPS	37		1	23.19			l	FOIA	R
1325403	m	400	3	12.00	tpl	i-n	l, l, a	FOIA, FOIA, FOIA	A, A, A
HdlPS	37		1	6.61			l	FOIA	A
1325603	f	400	3	12.00	t-l, t-r	n	a, t, t	FOIA, FOIA, FOIA	A, A, A
HdlPS	37		1	4.51			a	FBTC	2
1326103	f	400	3	12.00	t-r	s	a, r, r	FOA, UC, UC	A, 2, 2
HdlPS	34		1	22.39			r	UC	2
1327403	m	400	3	12.00	t-l	i-n	m, t, d	UC, UC, FBTC	A, 2, 2
HdlPS	50		1	9.27			d	UC	1
1328603	m	400	3	12.00	tml	s	t, l t	UC, FOA, UC	2, A, A
HdlPS	42		1	11.84			t	UC	A
1328803	m	400	3	12.00	f-r	i-n	c, l, l	UC, UC, UC	A, A, A
HdlPS	46		1	12.75			c	UC, UC	A
1328903	m	400	3	12.00	t-b	o	l, d, l	FOA, FOA, FOA	A, A, A
HdlPS	32		2	154.72			m, r	FOA, FOA	A, A
1330903	m	400	3	12.00	h-b, h-l	s	l, t, l	FOIA, FOIA, FOIA\	A, A, A\
HdlPS	36		5	111.07			l, d, t, t, l	FOIA, FOIA, FOIA, FOIA	A, A, A, A
								FOIA	A

Appendix K

MOEA's configuration details

The details about the Multiobjective Evolutionary Algorithm (MOEA) configuration, from chapter 5, are presented here.

Once every individual has been evaluated, parents are selected to reproduce and generate offspring. In this study, considering a population of N individuals, a group of $N/2$ parents are selected and are recombined until N offspring are produced. The offspring is then subjected to mutation (with a given probability, just as before for recombination), and a replacement strategy is put in place to select the N individuals that will make up the following generation. To do these steps, firstly, there is the need to rank the population, where non-dominated sorting was used.

For each individual, its rank is equal to the number of individuals dominating it plus one (e.g. all nondominated individuals are assigned rank 1). Afterwards, a fitness value was assigned by interpolating from the best individual (rank 1) to the worst (rank $n \leq N$) within each rank. In other words, this strategy iteratively searched for all the non-dominated solutions in the population that have not been labelled as belonging to a previous front. After labelling that new front, a front counter was incremented, and the process was repeated until all solutions were ranked.

Afterwards, the individuals within each rank were sorted according to the crowding distance, corresponding to the average side length of a cuboid defined by its nearest neighbours in the same front [Deb et al., 2002]. In essence, the larger the crowding distance is, the fewer solutions occupy the vicinity of a given individual. It is computed, for each objective m, in the following manner: sort individuals according to their fitness score f, assign an infinite value to boundary solutions (so that they are always selected), and compute the distance measure for the remaining solutions as given by Equation K.1. The overall crowding distance was calculated as the sum of the distance values concerning each objective.

$$F(i) = \frac{f(i+1)_m - f(i-1)_m)}{f_m^{max} - f_m^{min}}. \tag{K.1}$$

For parent selection, the population was ranked using non-dominated sorting, and then, within each rank, individuals were sorted by the crowding distance (the higher it was, the better rank they were assigned to). Parents were then chosen using binary tournaments and reproduced until N offspring were generated. Concerning the replacement strategy, there was used an elitist approach. Firstly, after evaluating the newly created offspring, the $2N$ individuals (current generation and offspring) were ranked with non-dominated sorting. Then, entire fronts were added into the new generation, starting from the rank 1 individuals, followed by rank 2, and so on, until a new set could no longer be accommodated. This last set of solutions was then sorted according to the crowding distance, and the better ones were chosen to fill out the rest of the new population.

Appendix L

MOEA's genotype-phenotype mapping example

A genotype-phenotype mapping example concerning the Multiobjective Evolutionary Algorithm (MOEA) from chapter 5 is presented here.

Suppose that for a 10-minute Seizure Prediction Horizon (SPH), a given individual is represented by three (not five for simplicity reasons) hyper-features, whose codification is presented in Table L.1.

The genotype-phenotype mapping transforms an individual's genotype into a set of hyper-features extractor that can perform sliding time-window analysis. The first three decoding steps consist in: i) finding the features that will be decoded to the phenotype; ii) constructing the hyper-features using the decoded features and the remaining genes; and iii) placing the hyper-features chronologically and adjusting the preictal period.

Concerning step i), for hyper-features A and B, the decoded features belong to the frequency domain due to their *active feature domain* gene value ("Frequency"). Hyper-feature A is decoded into a band relative power, namely alpha band due to the *active frequency feature* (band feature) and *active frequency band feature* (relative power). Hyper-feature B is decoded into SEF 50% due to the *active frequency feature* (spectral edge feature). Hyper-feature C is decoded into skewness, due to the *active feature domain* ("Time") and *active time feature* (statistical moment) gene values. This step is shown in L.1, more specifically in red.

Concerning step ii), the selected features, alpha band, SEF50%, and skewness, will now be collected for constructing hyper-features A, B, and C from electrodes Cz, O1 and T6, in windows of 15, 5 and 1 minutes, respectively. All possible window lengths are multiples of 5 seconds as the features were beforehand extracted in windows of 5 seconds. Then, to these windows, the correspondent mathematical operator is applied: the variance, the mean, and the integral for hyper-features A, B, and C, respectively.

Concerning step iii), each hyper-feature temporal position is obtained by adding

Table L.1: Example of an individual genotype. In this example, an individual is composed by only three hyper-features and not five, due to simplicity reasons. Each hyper-feature comprises twelve genes: active feature domain, active time feature, active frequency feature, active frequency band feature, statistical moment, hjorth parameter, relative band power, wavelet energy, spectral edge frequency, mathematical operator, electrode, window length, and delay.

Gene	A	B	C
Domain	Frequency	Frequency	Time
Time	Statistical moment	Hjorth parameter	Statistical moment
Frequency	Band division	Spectral edge	Band division
Frequency band feature	Relative power	Wavelet energy	Wavelet energy
Statistical moment	Kurtosis	Mean	Skewness
Hjorth parameter	Activity	Mobility	Complexity
Relative band power	Alpha	High gamma	Theta
Wavelet energy	D1	D7	A7
Spectral edge frequeny	75%	50%	90%
Mathematical operator	Variance	Mean	Integral
Electrode	Cz	O1	T6
Window length (minutes)	15	5	1
Delay (minutes)	30	10	20
Preictal Period (minutes)		30	

the correspondent delay to the preictal period gene. This gene allows analysing of a sequence of instants instead of only one instant. Additionally, it also allows adapting the preictal period duration, as it is determined by calculating the temporal distance from the first chronological hyper-feature to seizure onset. For a better understanding, this step is demonstrated in L.1, more specifically in orange.

Then, by setting the decoded hyper-features chronologically concerning the used preictal period, it is possible to perform label (preictal/interictal) and hyper-feature extraction through sliding-window analysis both in training and testing seizures for fitness function evaluation, as depicted in SL.2.

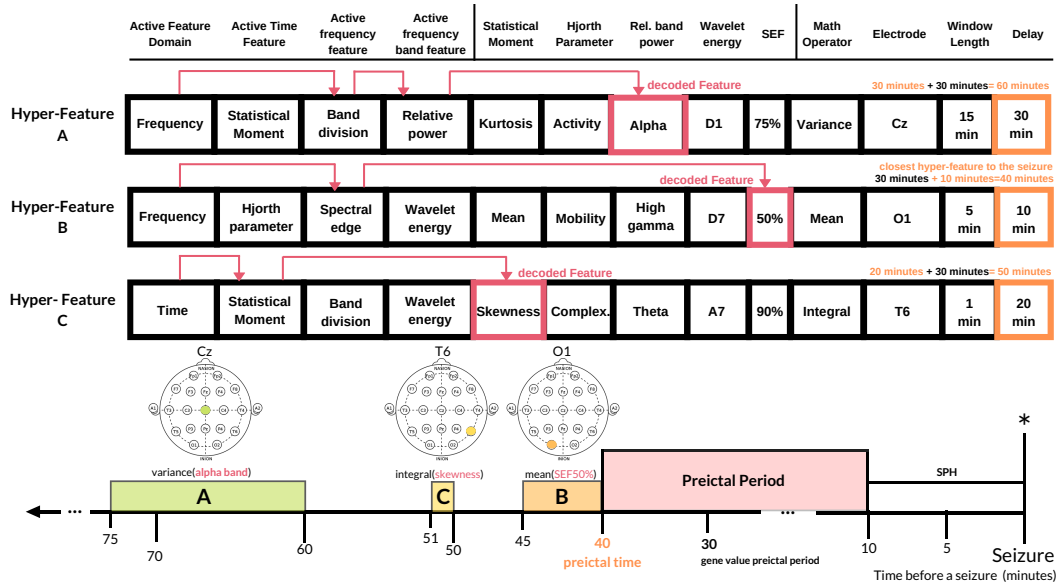


Figure L.1: The first three steps of decoding the genotype into the phenotype: i) decoding the dominant feature (in red); ii) constructing the hyper-features using the decoded feature and the remaining genes (in the timeline); iii) determining the hyper-features temporal position and finding the preictal period (in orange).

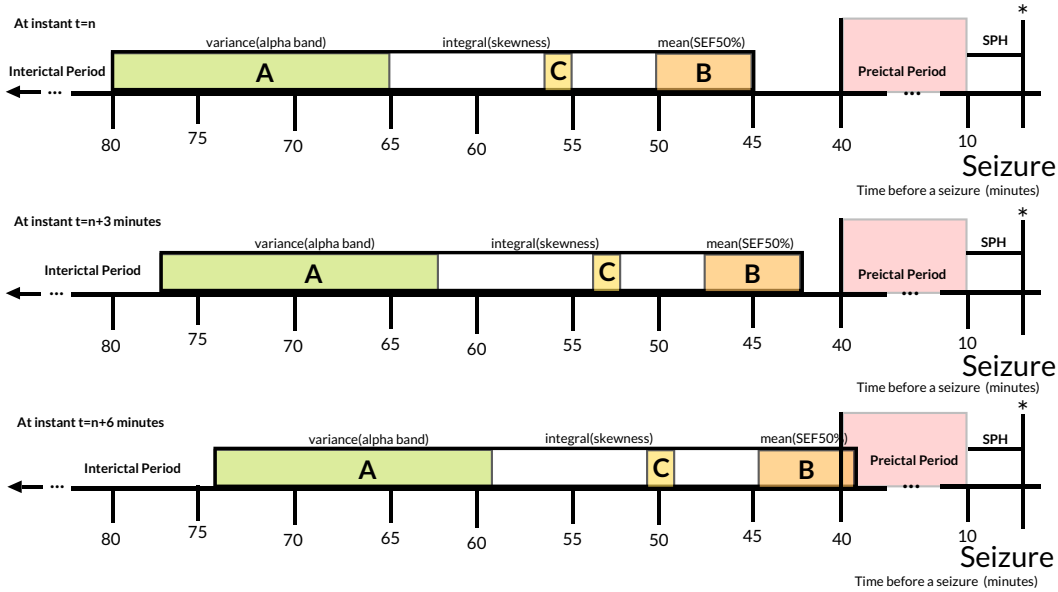


Figure L.2: With the hyper-features ordered chronologically, it is possible to perform feature extraction for a time-moving analysis and to label instants as belonging to the preictal/interictal period.

Appendix M

MOEA's evolution operators details and example

The mutation and the recombination operators concerning the Multiobjective Evolutionary Algorithm (MOEA), from chapter 5, are presented here.

M.1 Mutation operator details and example

Mutation, interpreted as a unitary step that will cause a random and unbiased change [Eiben and Smith, 2003], occurs in the following form for an individual: either one of the hyper-features or the preictal period gene is chosen randomly. When one of the hyper-features is chosen, one gene of that hyper-feature is then chosen randomly to mutate. If the preictal period gene is selected, its value will mutate. The remaining hyper-features and genes continue unaltered. Recombination is a stochastic operator that combines genetic information from two parents (individuals) into one or more offspring [Eiben and Smith, 2003].

The gene selection, despite random, is made by considering a weighted probability according to its neighbourhood: a gene's probability of being selected is proportional to its number of possible neighbourhood values. The higher the number of possible values (number of neighbours) for a gene, the higher its probability of being selected for mutation. Thus, the gene i probability selection $g(i)$ is computed as in (M.1), where G is the number of genes and $N_N(i)$ the gene respective number of neighbours:

$$g(i) = \frac{N_N(i)}{\sum_{j=1}^{G} N_N(j)}. \tag{M.1}$$

To perform a unary step, the mutation operator will act differently depending on the gene and its value since different genes have different neighbourhoods. One can understand all gene neighbourhoods as graphs: time instants, window-length and wave feature domains are graphs where the connected nodes have ordered val-

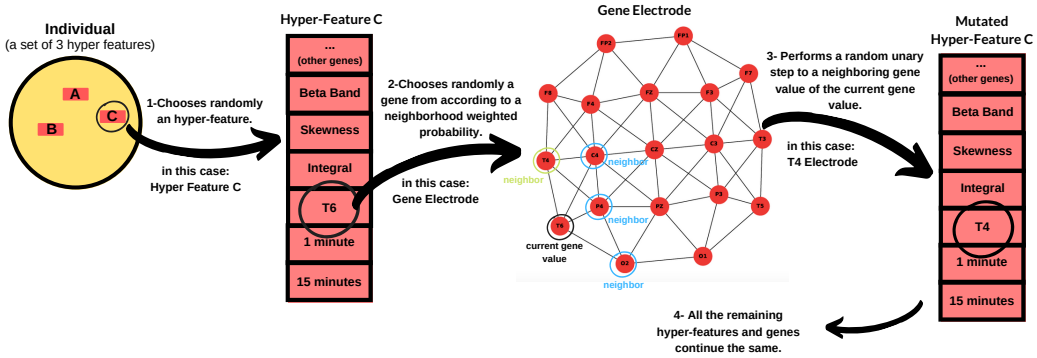


Figure M.1: How the mutation operator works in one individual.

ues. Mathematical operator and non-wave features are graphs where each node is connected to all nodes. Thus, a mutation can be interpreted as a random change from a gene node to a neighbouring one, randomly chosen.

This section provides an example using an individual. The mutation operator occurs in the following form, as illustrated in M.1: one of the hyper-features that composes the individual is chosen randomly (hyper-feature C), and then one gene of that hyper-feature is chosen randomly to mutate (electrode gene). Mutation will perform a random change from the current gene value node (T6) to one of its neighbours (T4, P4, C4, O2), in this case, to the T4 gene value. The remaining hyper-features and remaining genes from the mutated hyper-feature remain the same.

M.2 Recombination operator details and example

Recombination is a stochastic operator that combines genetic information from two parents (individuals) into one or more offspring [Eiben and Smith, 2003]. After selecting two parents to reproduce, this operator performs the recombination of all paired hyper-features. Thus, hyper-feature pairing is the first step and is performed by calculating their Manhattan distances and matching the closest ones. The distance between two hyper-features equals the number of steps needed to go from one value to another by taking the shortest path. By representing $D(f_a, f_i)$ as the distance between the hyper-features f_a and f_i where a is the index of the fixed feature from one parent and i the feature index that iterates the other parent F number of features, the matched feature index m will be given by equation (M.2). Thus, features f_m and f_a match.

$$m = \arg\min_i D(f_a, f_i), 1 \leq i \leq F. \tag{M.2}$$

$D(f_a, f_i)$ can be described as the summation of all gene Manhattan distances, where G is the number of genes from an feature and $d(f_{a_i}, f_{b_i})$ is the distance between two i gene values from features f_a and f_b, presented in equation (M.3).

$$D(f_a, f_b) = \sum_{i=1}^{G} d(f_{a_i}, f_{b_i}), 1 \leq i \leq G. \tag{M.3}$$

After the hyper-feature matching operation finishes, the recombination operator works at the hyper-feature gene level. Consequently, each offspring gene value was obtained by choosing a random node belonging to the shortest path between the correspondent two-parent gene values. The distance between two gene values equals the number of steps needed to go from one value to another by taking the shortest path. Due to computational simplifications, the feature matching step did not consider the total distance of the minimum components as it used a greedy approach. More specifically, one of the parents was selected, and for each feature in his genotype, the closest feature from the other parent was matched. This was done iteratively, which means that if a feature was already matched with a previous component, it could not be used, even if a lower $D(f_a, f_b)$ was obtained with the new matching.

This section provides an example between two hyper-features concerning recombination at the hyper-feature gene level. Thus, in this stage, hyper-features from both parents were already paired. M.2 illustrates the recombination operation concerning one of the paired hyper-features (orange and red), where each gene is recombined (the same process is then repeated with all paired hyper-features). The new gene value is a random node between the shortest path of the two-parent gene nodes. When several possible paths are possible, as in the case of the electrodes, one of the shortest paths is, beforehand, randomly chosen. The recombined hyper-feature is presented in green.

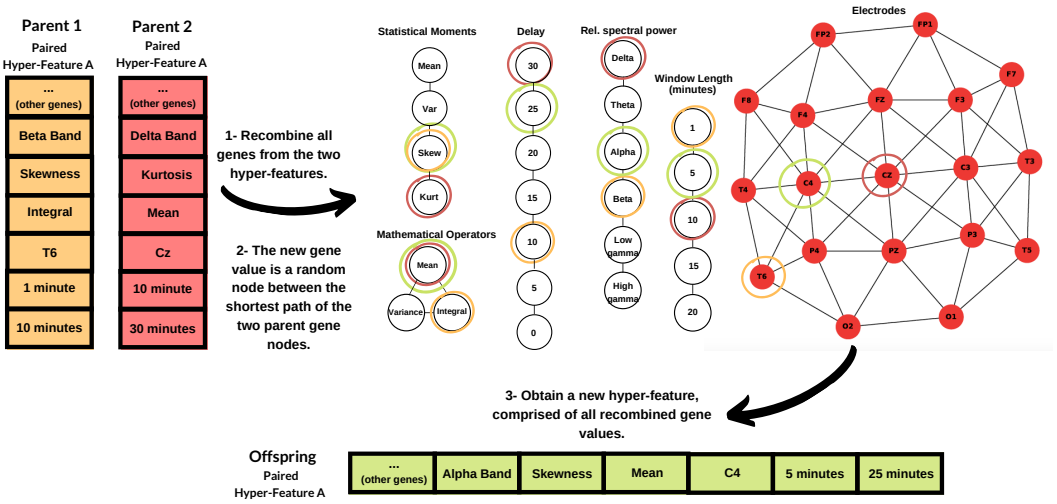


Figure M.2: An example of how the recombination operator work at the hyper-feature gene level. The genotype of the parents' hyper-features is presented in orange and red, while green represents the recombined hyper-feature.

Appendix N

MOEA's neighbourhood details

The details concerning the Multiobjective Evolutionary Algorithm (MOEA) established neighbourhoods, from chapter 5, are presented here.

Window length and time instant genes have a straightforward ordering, as their values correspond to increasing/decreasing discrete time intervals. Electrodes' neighbourhood is based on their scalp position. As there is no decreasing/increasing relationship among the mathematical operators (mean, median, variance, integral), all genes were considered neighbours of each other.

However, the extracted features can be divided into several groups and subgroups, as seen in Fig. 3.b). Firstly, they can be divided concerning their time/frequency domain. Then, time-domain features can be divided into statistical moments (ordered by their n-th moment: mean, variance, skewness, kurtosis) or into Hjorth Parameters (ordered by their n-th moment: activity, mobility, complexity).

Frequency-domain features can be divided into spectral edge features (ordered by the frequency: 50%, 75%, and 90%) or frequency band ones. Frequency band features can be divided into relative spectral power bands (ordered by their frequency range: delta, theta, beta, alpha, low-gamma, high-gamma) and the energy from wavelet decomposition levels (ordered by their decomposition levels: D1, D2, D3, D4, D5, D6, D7, and D8). Please note that the used overall division might lead to discussion, as wavelet transforms may be considered forms of time-frequency representations and not solely belonging to the frequency domain. Nevertheless, it was believed that these features are conceptually similar to relative power spectral bands.

Appendix O

Mathematical formulation of the MOEA phenotype study

The mathematical formulation of the Multiobjective Evolutionary Algorithm (MOEA) phenotype study, from chapter 5, is presented here.

As Evolutionary Algorithmss (EAs) are associated with random components (in this case: initialisation, parent selection, and evolution operators), it is possible to obtain, for each execution, a different solution (set of hyper-features) with similar performance. Thus, performing a phenotype study aims to understand the overall influence of each gene value concerning the obtained hyper-features.

For a given hyper-feature j, a simple approach can study each decoded gene individually. It is possible to calculate the gene presence using Equation (O.1) by assigning to a binary value (1/0) considering the gene value presence in a hyper-feature.

$$presence(gene, value) = \begin{cases} 1, & \text{if } gene = value. \\ 0, & \text{otherwise.} \end{cases} . \tag{O.1}$$

Then, by applying the previous equation for all F hyper-features that compose an individual, one obtains the correspondent gene value presence for an individual. After this, one can compute the correspondent normalised gene value presence $Presence(value)$ for all individuals I, as demonstrated in Equation (O.2), respectively. As 30 executions were performed, and as each individual is composed of 5 hyper-features, $I = 30$, $F = 5$, and G gene values (which depends on the gene):

$$Presence(value) = \frac{1}{I} \sum_{i=1}^{I} \max \left\{ presence(gene_{ijk}, value) | j = 1, \ldots, F; k = 1, \ldots, G \right\} . \tag{O.2}$$

Appendix P

MOEA study's control method

The developed control methodology that was developed to evaluate the performance of the Multiobjective Evolutionary Algorithm (MOEA), from chapter 5, is presented here.

The existing clinical trial inspired this control method that deals with EEG seizure susceptibility detection, namely the Seizure Advisory System Feasibility Study (NCT01043406) [Cook et al., 2013] and other previous works, such as [Direito et al., 2017]. Its development was merely inspired and not fully reproduced to simplify the process.

The Neurovista Advisory System used a patient-specific layered structure, including filtering, feature extraction, and classification. Input signals were filtered by octave-wide digital filters, from 8Hz to 128Hz range. Notch filters were also optionally available. The signals were firstly segmented where they extracted the average energy, Teager-Kaiser energy, and line-length into windows of 5 seconds, and, secondly, these features were normalised. This output resulted in 288 features, as these resulted from the combination of 16 available Invasive Electroencephalogram (iEEG) input channels, 6 filter/normalisation options, and 3 available features. The 16 best performing features are selected during training using a backward elimination feature selection method based on Hilbert-Schmidt Independence Criterion (BAHSIC). The selected features were used to train a classification model inspired by two types of classification models, decision trees and a k-nearest neighbours (kNN). A given feature vector was classified by using several decision surfaces.

Simply put, the input was first compared to the first decision surface to determine the binary side of the surface where the point was on. Then, a new decision surface is selected to perform a second surface comparison depending on the previous binary answer. This process continues until reaching a total of ten decision layers. Thus, the feature space was divided into 2^{10} partitions. The used surfaces were chosen during training. To each, it was assigned a relative measure of seizure risk concerning their time distance to a seizure event. The training phase used a hold-out or cross-fold validation method to get the required configuration parameters. Then, the classifier

output was filtered and thresholded to prevent rapid changes and, thus, be robust to noise.

This work's extracted features replaced the advisory system's filtering/decomposition and feature extraction processes. This way, the same data was used for the MOEA and control method. Then, the best set was selected from all possible combinations of features and electrodes. Feature selection was based on two phases: using two filter methods [Dash and Liu, 1997] and an embedded one. Concerning the filter methods: firstly, there were selected the best 100 features with Pearson's linear correlation coefficient and secondly, the best 25 with Receiver Operating Characteristic (ROC)'s Area Under the Curve (AUC). Then, the wrapper method Recursive Feature Elimination (RFE) was used to select an optimal set of features. The filter method had the objective of substantially decreasing the number of features and thus computational power, as it would take the RFE a substantially long period to compute.

There are two important configuration options when using the RFE: the used algorithm to help select the features (the estimator) and the number of features. For the estimator, there was used a linear Support Vector Machines (SVM) [Direito et al., 2011]. For the number of features, the first three chronological seizures were used for a grid-search (3, 5, 7, 10, 15, 20 features) that optimises the Sample Sensitivity (S_{ss}) and Sample Specificity (S_{sp}): $\sqrt{S_{ss} * S_{sp}}$ (also named here as fitness).

The grid search was used to find the best preictal period (30, 35, 40, 45, 50, 55, 60 minutes). For the grid search, there was used a 3-fold as in [Direito et al., 2017]: fold 1 used seizures #1 and #2 for training, and #3 for validation; fold 2 used seizures #1 and #3 for training, and #2 for validation; and fold 3 used seizures #2 and #3 for training, and #1 for validation. The remaining seizures comprised the testing group. Please note that there was used the RFE with the SVM estimator instead of a backward feature elimination based on the BAHSIC criterion, as its code implementation was more easily available on Python scikit-learn library.

The classification model was adapted: a Random Forest was used since the one from the advisory system consists of an ensemble of many individual decision tree classifiers. Each decision tree is trained with different data distributions. Thus, it was intended to mimic the ten layer structure along with the different decision surfaces by using the random forest algorithm. In testing, it was also implemented the Firing Power to smooth the output over time, with a similar threshold value to the MOEA approach (0.7). Refractory periods were also used. This methodology is patient-specific.

Appendix Q

MOEA's study full results

The control method and Multiobjective Evolutionary Algorithm (MOEA) seizure prediction results for all patients, from chapter 5, are presented here.

Table Q.1: The results for all patients from the MOEA study.

ID	Training			Testing		Control Method		
	Preictal (min)	Fitness (SS/SP/Comf.)	# Pareto Front Ind.	SS (0-1) FPR/h	Surrogate SS (0-1)	SS (0-1) FPR/h	Surrogate SS (0-1)	Above chance
102	**48.44±9.32**	**0.96±0.05 0.94±0.06 60.60±0.18**	**1671**	**0.21±0.12 0.24±0.12**	**0.19±0.05**	**0.13 0.09**	**0.07±0.26**	**MOEA**
202	**51.41±7.14**	**0.98±0.03 0.98±0.03 0.73±0.16**	**2569**	**0.00±0.06 0.14±0.09**	**0.12±0.09**	**1.00 0.00**	**0.07±0.25**	**Control**
500	51.53±9.60	0.98±0.02 0.99±0.01 0.63±0.18	2354	0.12±0.15 0.18±0.06	0.15±0.06	0.50 0.28	0.38±0.48	-
1200	50.02±7.52	0.95±0.05 0.95±0.05 0.57±0.19	1092	0.08±0.26 0.24±0.22	0.16±0.15	0.00 0.33	0.20±0.40	-
1500	49.46±8.08	0.98±0.02 0.99±0.02 0.71±0.18	2503	0.03±0.18 0.35±0.32	0.22±0.13	0.00 0.00	0.00±0.00	-
1600	**50.0±7.63**	**0.96±0.04 0.97±0.03 0.67±0.19**	**1801**	**0.12±0.32 0.04±0.07**	**0.03±0.06**	**0.00 1.19**	**0.43±0.50**	**MOEA**
2100	47.13±8.26	0.95±0.05 0.95±0.06 0.70±0.18	1452	0.19±0.19 0.24±0.11	0.18±0.08	0.00 0.05	0.03±0.18	-
2300	50.65±7.2	0.96±0.04 0.96±0.02 0.74±0.13	1558	0.06±0.24 0.30±0.30	0.19±0.12	0.00 0.38	0.53±0.50	-
3300	**50.86±9.44**	**0.96±0.04 0.96±0.03 0.69±0.16**	**1908**	**0.23±0.42 0.26±0.17**	**0.20±0.12**	**1.00 0.00**	**0.17±0.37**	**MOEA Control**
4500	50.30±9.63	0.97±0.04 0.96±0.04 0.75±0.14	1727	0.18±0.38 0.33±0.16	0.24±0.11	0.00 0.69	0.43±0.50	-
5800	45.84±6.52	0.97±0.04 0.97±0.03 0.72±0.15	2090	0.12±0.21 0.35±0.15	0.24±0.13	0.00 0.21	0.15±0.36	-
6000	**52.99±9.36**	**0.97±0.03 0.96±0.03 0.73±0.15**	**1900**	**0.13±0.34 0.22±0.11**	**0.19±0.10**	**1.00 0.00**	**0.23±0.42**	**Control**
7200	50.38±8.5	0.97±0.03 0.97±0.03 0.72±0.15	2207	0.04±0.21 0.54±0.39	0.31±0.20	0.00 0.00	0.00±0.00	-
8100	**47.04±7.48**	**0.93±0.06 0.90±0.06 0.68±0.16**	**937**	**0.18±0.27 0.18±0.06**	**0.14±0.06**	**0.00 0.06**	**0.02±0.13**	**MOEA**
8902	**51.17±9.08**	**0.95±0.05 0.95±0.06 0.77±0.11**	**1154**	**0.20±0.27 0.16±0.09**	**0.24±0.08**	**1.00 0.00**	**0.05±0.22**	**Control**

11002	52.94±10.65	0.98±0.04 0.98±0.03 0.69±0.19	2507	0.00±0.03 0.27±0.19	0.20±0.13	0.00 0.31	0.17±0.37	-
11502	48.17±9.45	0.92±0.06 0.89±0.03 0.71±0.14	930	0.06±0.06 0.14±0.04	0.15±0.04	0.00 1.20	0.31±0.46	-
12702	**53.8±7.61**	**0.99±0.02 0.99±0.02 0.86±0.15**	**2499**	**0.16±0.17 0.21±0.06**	**0.16±0.05**	**0.33 0.43**	**0.18±0.38**	**Control**
21602	49.91±9.55	0.97±0.03 0.97±0.03 0.70±0.16	1896	0.16±0.14 0.32±0.09	0.27±0.06	0.00 0.21	0.08±0.28	-
21902	**51.40±9.55**	**0.95±0.04 0.95±0.04 0.67±0.19**	**1457**	**0.00±0.07 0.18±0.12**	**0.13±0.08**	**1.00 0.00**	**0.20±0.40**	**Control**
22602	**51.74±7.82**	**0.98±0.02 0.96±0.03 0.75±0.14**	**2030**	**0.17±0.26 0.10±0.08**	**0.17±0.07**	**0.50 0.00**	**0.10±0.30**	**Control**
23902	50.33±10.39	0.97±0.03 0.96±0.03 0.67±0.17	2120	0.08±0.18 0.22±0.09	0.23±0.13	0.00 3.43	0.43±0.50	-
26102	**49.11±7.34**	**0.91±0.06 0.91±0.05 0.63±0.15**	**684**	**0.13±0.33 0.34±0.11**	**0.24±0.10**	**1.00 0.00**	**0.17±0.37**	**Control**
30802	**51.73±8.31**	**0.98±0.03 0.97±0.05 0.64±0.17**	**1919**	**0.19±0.22 0.19±0.07**	**0.20±0.06**	**0.60 0.32**	**0.34±0.48**	**Control**
32502	**53.14±8.70**	**0.98±0.03 0.98±0.02 0.76±0.14**	**2570**	**0.12±0.18 0.35±0.13**	**0.27±0.08**	**0.33 0.91**	**0.17±0.37**	**Control**
46702	**51.05±8.08**	**0.97±0.03 0.96±0.03 0.72±0.17**	**1794**	**0.25±0.31 0.28±0.19**	**0.19±0.12**	**0.50 0.67**	**0.45±0.50**	**MOEA**
50802	**50.36±7.44**	**0.98±0.03 0.98±0.02 0.72±0.17**	**2221**	**0.03±0.13 0.22±0.09**	**0.15±0.07**	**0.50 1.45**	**0.20±0.40**	**Control**
51002	51.75±7.26	0.92±0.07 0.92±0.07 0.76±0.13	575	0.08±0.10 0.18±0.09	0.16±0.11	0.60 1.84	0.68±0.47	-
52302	51.0±8.17	0.98±0.03 0.98±0.02 0.75±0.14	2499	0.01±0.08 0.24±0.18	0.16±0.10	0.00 0.00	0.00±0.00	-
53402	**53.73±7.61**	**0.98±0.03 0.97±0.03 0.71±0.14**	**2447**	**0.30±0.33 0.26±0.06**	**0.20±0.06**	**0.50 0.03**	**0.02±0.13**	**MOEA Control**
55202	53.26±9.34	0.95±0.05 0.95±0.04 0.69±0.17	1352	0.18±0.16 0.23±0.06	0.21±0.05	0.00 0.08	0.12±0.32	-
56402	**53.23±6.88**	**0.98±0.03 0.96±0.04 0.75±0.16**	**2105**	**0.05±0.21 0.14±0.10**	**0.12±0.10**	**1.00 0.36**	**0.27±0.44**	**Control**
58602	52.35±11.14	0.97±0.03 0.97±0.03 0.72±0.15	1369	0.05±0.12 0.25±0.10	0.23±0.11	0.00 0.49	0.21±0.41	-
63502	48.15±9.05	0.98±0.03 0.97±0.03 0.64±0.18	2374	0.10±0.21 0.14±0.06	0.13±0.07	0.50 2.43	0.48±0.50	-
64702	54.85±7.57	0.97±0.03 0.96±0.03 0.70±0.18	2008	0.08±0.19 0.14±0.07	0.13±0.08	0.50 0.76	0.38±0.49	-
70302	**46.97±8.51**	**0.96±0.04 0.93±0.06 0.64±0.15**	**570**	**0.13±0.10 0.26±0.06**	**0.18±0.06**	**0.43 0.25**	**0.19±0.39**	**Control**
70902	**48.34±7.79**	**0.98±0.03 0.98±0.02 0.71±0.20**	**1930**	**0.07±0.17 0.21±0.10**	**0.13±0.07**	**0.50 0.88**	**0.28±0.45**	**Control**
71102	**51.83±6.08**	**0.97±0.03 0.97±0.03 0.74±0.19**	**2090**	**0.11±0.21 0.19±0.12**	**0.18±0.08**	**1.00 1.22**	**0.62±0.49**	**Control**
71502	47.62±7.58	0.95±0.05 0.93±0.06 0.62±0.16	1451	0.05±0.14 0.16±0.07	0.14±0.06	0.50 10.14	0.33±0.47	-
71802	**51.71±8.48**	**0.95±0.04 0.96±0.04 0.70±0.20**	**1164**	**0.12±0.21 0.21±0.08**	**0.17±0.07**	**0.50 10.63**	**0.30±0.46**	**Control**
73002	**51.07±9.61**	**0.98±0.02 0.99±0.02 0.65±0.18**	**2560**	**0.04±0.14 0.17±0.08**	**0.14±0.08**	**1.00 1.30**	**0.20±0.40**	**Control**

75102	**50.25±8.93**	**0.97±0.03 0.97±0.03 0.69±0.18**	**1924**	**0.19±0.26 0.24±0.09**	**0.20±0.08**	**0.50 0.66**	**0.23±0.42**	**Control**
75202	**49.45±8.64**	**0.98±0.02 0.98±0.02 0.65±0.18**	**2225**	**0.19±0.17 0.14±0.05**	**0.15±0.06**	**0.00 0.17**	**0.07±0.25**	**MOEA**
79502	**47.62±10.08**	**0.94±0.06 0.87±0.06 0.69±0.17**	**1231**	**0.21±0.17 0.18±0.08**	**0.15±0.07**	**0.25 0.87**	**0.22±0.41**	**MOEA**
80602	52.12±9.73	0.98±0.02 0.99±0.02 0.76±0.15	2230	0.04±0.13 0.12±0.04	0.10±0.05	0.00 0.14	0.03±0.18	-
80702	50.65±8.26	0.96±0.04 0.95±0.05 0.73±0.18	1580	0.10±0.20 0.14±0.08	0.15±0.07	0.00 0.00	0.00±0.00	-
81402	**50.58±8.10**	**0.98±0.02 0.99±0.02 0.80±0.11**	**2110**	**0.41±0.26 0.19±0.05**	**0.16±0.05**	**0.33 0.00**	**0.08±0.27**	**MOEA Control**
85202	**52.02±11.45**	**0.96±0.04 0.96±0.04 0.67±0.18**	**1389**	**0.42±0.40 0.25±0.13**	**0.21±0.09**	**0.00 3.82**	**0.32±0.46**	**MOEA**
92102	**48.98±7.10**	**0.97±0.03 0.96±0.03 0.70±0.16**	**1595**	**0.40±0.32 0.27±0.07**	**0.18±0.07**	**0.00 2.37**	**0.37±0.48**	**MOEA**
93402	50.31±8.29	0.99±0.02 0.99±0.02 0.82±0.10	2524	0.11±0.21 0.32±0.08	0.25±0.10	0.00 0.00	0.00±0.00	-
93902	**48.66±7.89**	**0.97±0.03 0.95±0.04 0.66±0.20**	**1852**	**0.37±0.28 0.23±0.15**	**0.20±0.10**	**0.00 2.28**	**0.47±0.50**	**MOEA**
94402	48.76±9.0	0.98±0.03 0.96±0.03 0.74±0.15	1748	0.13±0.16 0.36±0.13	0.24±0.07	0.00 1.42	0.48±0.50	-
95202	50.41±10.48	0.96±0.03 0.95±0.03 0.66±0.17	1144	0.09±0.14 0.16±0.06	0.13±0.06	0.00 0.02	0.00±0.00	-
96002	50.72±7.60	0.98±0.02 0.99±0.02 0.81±0.16	2725	0.16±0.18 0.22±0.06	0.16±0.05	0.00 0.30	0.14±0.35	-
98102	**51.01±8.16**	**0.97±0.04 0.93±0.05 0.65±0.19**	**1194**	**0.32±0.28 0.11±0.07**	**0.11±0.06**	**0.50 2.46**	**0.22±0.41**	**MOEA Control**
100002	51.65±9.36	0.98±0.03 0.98±0.03 0.73±0.13	2472	0.13±0.10 0.27±0.06	0.18±0.05	0.00 0.15	0.10±0.30	-
101702	**49.54±8.12**	**0.96±0.04 0.96±0.04 0.64±0.18**	**1448**	**0.34±0.31 0.24±0.14**	**0.20±0.10**	**0.00 0.15**	**0.10±0.30**	**MOEA**
102202	**50.09±8.40**	**0.96±0.05 0.94±0.05 0.68±0.15**	**1544**	**0.22±0.22 0.18±0.06**	**0.15±0.06**	**0.00 0.06**	**0.03±0.16**	**MOEA**
103002	**54.48±9.39**	**0.98±0.03 0.98±0.02 0.70±0.14**	**2384**	**0.17±0.22 0.13±0.04**	**0.12±0.05**	**0.67 1.20**	**0.50±0.50**	**MOEA**
103802	**50.36±8.95**	**0.97±0.03 0.97±0.03 0.76±0.15**	**2084**	**0.26±0.24 0.18±0.11**	**0.16±0.07**	**0.00 0.06**	**0.00±0.00**	**MOEA**
104602	**52.63±8.68**	**0.97±0.04 0.98±0.03 0.74±0.16**	**2071**	**0.33±0.29 0.26±0.16**	**0.23±0.09**	**0.50 0.00**	**0.10±0.30**	**MOEA Control**
109202	52.89±9.84	0.95±0.03 0.95±0.04 0.72±0.14	1391	0.06±0.24 0.34±0.19	0.25±0.15	0.00 0.00	0.00±0.00	-
109502	54.82±8.58	0.98±0.02 0.99±0.02 0.80±0.14	2405	0.11±0.31 0.14±0.08	0.13±0.08	0.00 0.30	0.03±0.18	-
109602	**53.16±7.93**	**0.99±0.01 0.99±0.01 0.86±0.11**	**2373**	**0.26±0.44 0.16±0.06**	**0.14±0.08**	**1.00 1.54**	**0.47±0.50**	**MOEA Control**
110602	**52.32±8.01**	**0.98±0.03 0.98±0.03 0.74±0.14**	**2256**	**0.37±0.29 0.20±0.08**	**0.19±0.07**	**0.00 0.00**	**0.47±0.00**	**MOEA**
111902	**49.25±9.42**	**0.95±0.06 0.93±0.06 0.59±0.19**	**1361**	**0.08±0.19 0.12±0.13**	**0.13±0.11**	**0.50 0.00**	**0.07±0.25**	**Control**
112402	51.96±9.37	0.96±0.05 0.96±0.05 0.68±0.13	1168	0.16±0.12 0.25±0.09	0.19±0.06	0.22 0.32	0.20±0.40	-

113902	**52.06±7.22**	**0.97±0.03 0.97±0.03 0.74±0.13**	**2229**	**0.28±0.12 0.07±0.06**	**0.07±0.05**	**0.00 0.00**	**0.00±0.00**	**MOEA**
114702	**47.77±7.98**	**0.98±0.02 0.98±0.03 0.76±0.21**	**2541**	**0.16±0.16 0.35±0.12**	**0.24±0.08**	**0.40 0.60**	**0.17±0.38**	**Control**
114902	**52.95±8.07**	**0.98±0.03 0.98±0.03 0.77±0.15**	**2566**	**0.33±0.21 0.10±0.04**	**0.09±0.04**	**0.25 0.55**	**0.16±0.36**	**MOEA**
115202	**50.25±9.16**	**0.96±0.03 0.97±0.04 0.80±0.19**	**1610**	**0.25±0.18 0.22±0.08**	**0.19±0.06**	**0.00 0.17**	**0.06±0.23**	**MOEA**
123902	51.17±9.63	0.98±0.02 0.99±0.02 0.71±0.15	2521	0.07±0.18 0.14±0.08	0.11±0.08	0.50 0.25	0.33±0.47	-
125002	51.5±10.05	0.97±0.03 0.97±0.03 0.64±0.16	1622	0.14±0.15 0.20±0.09	0.17±0.07	0.25 1.80	0.27±0.44	-
1235103	51.25±8.43	0.98±0.03 0.98±0.02 0.71±0.16	2226	0.09±0.22 0.13±0.08	0.11±0.08	0.50 7.91	0.75±0.43	-
1307103	51.08±9.03	0.97±0.05 0.95±0.06 0.71±0.14	1998	0.10±0.12 0.11±0.06	0.16±0.04	0.20 0.13	0.33±0.47	-
1308503	**52.06±9.81**	**0.98±0.02 0.99±0.02 0.75±0.12**	**2619**	**0.26±0.44 0.20±0.13**	**0.15±0.11**	**1.00 2.39**	**0.43±0.50**	**MOEA Control**
1308603	**50.32±8.75**	**0.98±0.03 0.98±0.03 0.73±0.17**	**2093**	**0.41±0.49 0.19±0.09**	**0.15±0.08**	**0.00 0.36**	**0.20±0.40**	**MOEA**
1310803	**50.51±7.93**	**0.95±0.06 0.94±0.06 0.61±0.15**	**1770**	**0.01±0.06 0.20±0.07**	**0.17±0.10**	**1.00 0.00**	**0.10±0.30**	**Control**
1312903	**50.31±7.82**	**0.98±0.02 0.98±0.02 0.66±0.18**	**2363**	**0.12±0.32 0.12±0.09**	**0.09±0.08**	**0.00 0.33**	**0.37±0.48**	**MOEA**
1315203	53.08±7.29	0.98±0.03 0.97±0.04 0.67±0.17	2072	0.06±0.24 0.13±0.07	0.11±0.08	0.00 0.00	0.00±0.00	-
1319203	**50.42±6.78**	**0.95±0.05 0.95±0.05 0.64±0.20**	**1385**	**0.01±0.06 0.20±0.07**	**0.09±0.08**	**0.50 0.23**	**0.28±0.45**	**Control**
1321803	**49.79±8.43**	**0.97±0.03 0.97±0.03 0.67±0.19**	**2522**	**0.54±0.50 0.14±0.26**	**0.18±0.14**	**1.00 0.38**	**0.23±0.42**	**MOEA Control**
1322803	**51.59±8.8**	**0.98±0.03 0.97±0.03 0.74±0.17**	**2471**	**0.09±0.28 0.24±0.12**	**0.19±0.11**	**1.00 0.00**	**0.10±0.30**	**Control**
1324903	50.95±8.00	0.91±0.06 0.95±0.05 0.66±0.15	880	0.25±0.30 0.22±0.10	0.23±0.08	0.00 1.63	0.28±0.45	-
1325103	51.53±8.27	0.96±0.04 0.92±0.05 0.70±0.19	1275	0.09±0.22 0.13±0.07	0.11±0.08	0.00 0.30	0.10±0.30	-
1325403	**48.62±8.17**	**0.97±0.03 0.98±0.02 0.73±0.17**	**2197**	**0.21±0.41 0.28±0.29**	**0.18±0.16**	**0.00 0.00**	**0.00±0.00**	**MOEA**
1325603	**48.41±8.99**	**0.94±0.05 0.95±0.05 0.66±0.21**	**1190**	**0.12±0.32 0.28±0.29**	**0.20±0.22**	**1.00 40.00**	**0.60±0.49**	**Control**
1326103	49.25±8.70	0.97±0.03 0.96±0.03 0.77±0.16	1510	0.09±0.28 0.30±0.13	0.21±0.10	0.00 0.00	0.00±0.00	-
1327403	50.25±9.42	0.95±0.05 0.95±0.04 0.62±0.16	1001	0.00±0.00 0.10±0.10	0.09±0.10	0.00 0.34	0.20±0.40	-
1328603	**50.18±9.70**	**0.95±0.06 0.93±0.06 0.67±0.16**	**1334**	**0.16±0.37 0.33±0.13**	**0.24±0.11**	**1.00 1.90**	**0.63±0.48**	**Control**
1328803	52.76±8.97	0.99±0.02 0.99±0.02 0.79±0.12	2658	0.04±0.20 0.20±0.24	0.13±0.15	0.00 0.69	0.10±0.30	-
1328903	**50.15±9.77**	**0.97±0.03 0.98±0.02 0.71±0.20**	**2304**	**0.23±0.28 0.20±0.05**	**0.15±0.07**	**0.00 1.21**	**0.23±0.42**	**MOEA**
1330903	**46.88±8.63**	**0.98±0.03 0.98±0.03 0.67±0.19**	**2076**	**0.19±0.15 0.13±0.04**	**0.08±0.03**	**0.20 0.07**	**0.08±0.27**	**MOEA Control**

Appendix R

MOEA study's preictal period discussion

Some examples of the preictal periods' histograms concerning chapter 5 are presented here. More specifically, patients 21602, 21902, 30802, and 32502 are presented, which had, as mean preictal period the following values, respectively: 49.91±9.55, 51.40±9.55, 51.73±8.31, and 53.14±8.70 minutes. In the GitHub page of the paper, one can find these histograms for all patients. The most frequent preictal period (the mode) is also presented for each patient. The Multiobjective Evolutionary Algorithm (MOEA) can find solutions for many possible preictal periods in all patients.

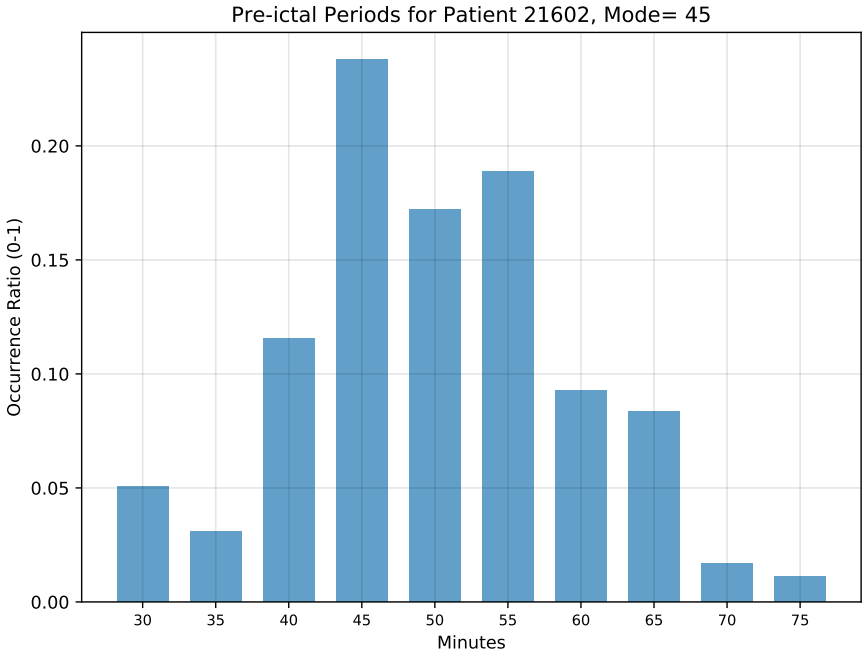


Figure R.1: The obtained preictal periods, for all MOEA solutions, for patient 21602.

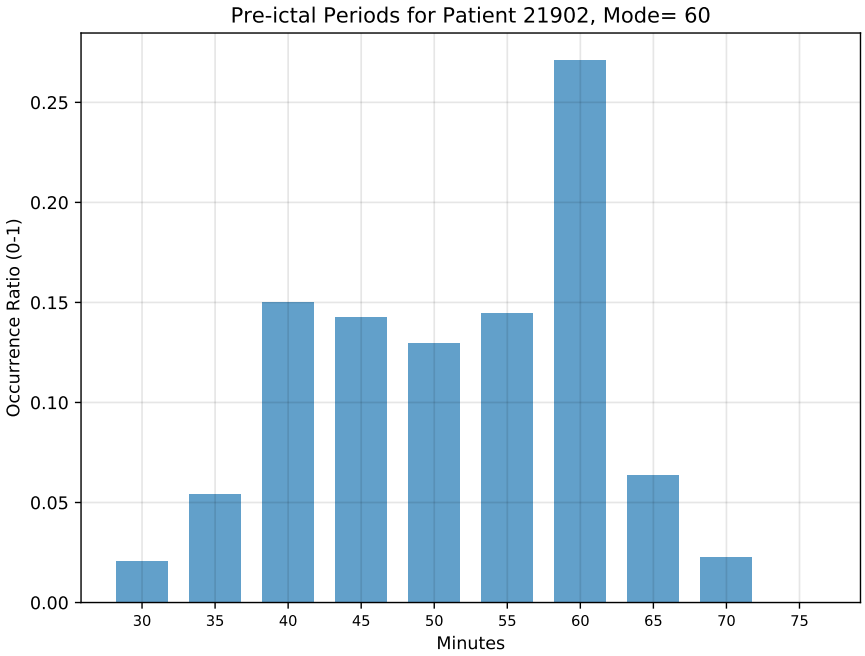


Figure R.2: The obtained preictal periods, for all MOEA solutions, for patient 21902.

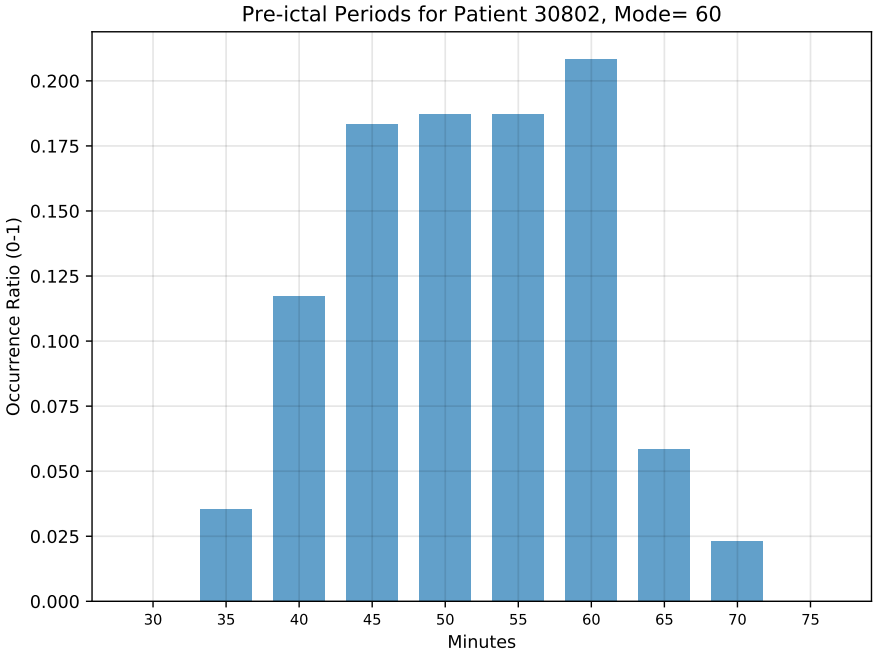


Figure R.3: The obtained preictal periods, for all MOEA solutions, for patient 30802.

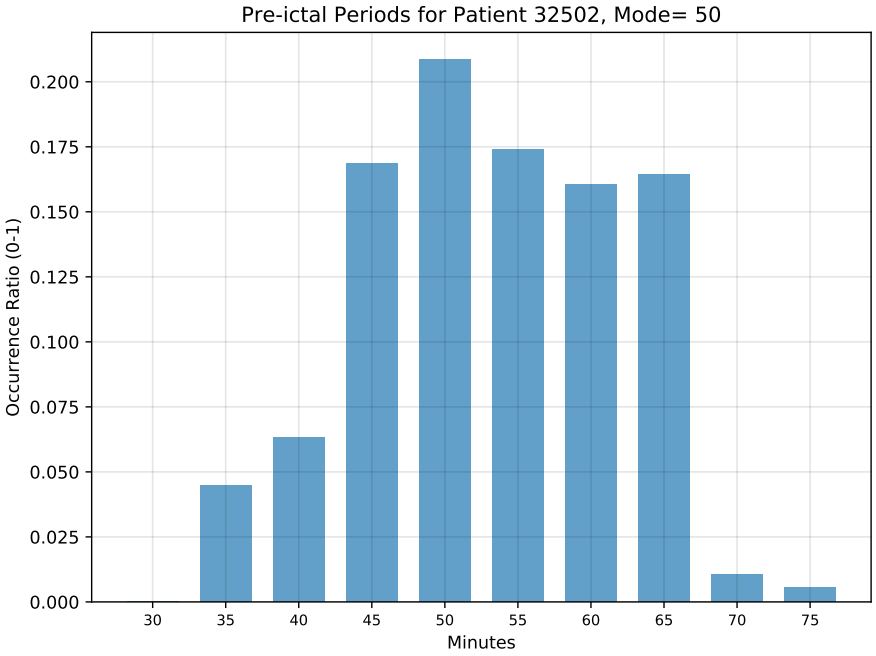


Figure R.4: The obtained preictal periods, for all MOEA solutions, for patient 32502.

Appendix S

MOEA study's electrodes and lobes discussion

Some examples of the number of different electrodes and lobes histograms concerning chapter 5 are presented here. More specifically, patients 1200, 12702, 55202, 81402, and 1319203 are presented, which had, as most common set (mode) of electrodes and lobes the following values, respectively: 4 and 3; 2 and 1; 3 and 2; 2 and 2; and 4 and 2. On the GitHub page of this paper, one can find these histograms for all patients. These findings evidence the relevance of the patient comfort metric in the Multiobjective Evolutionary Algorithm (MOEA).

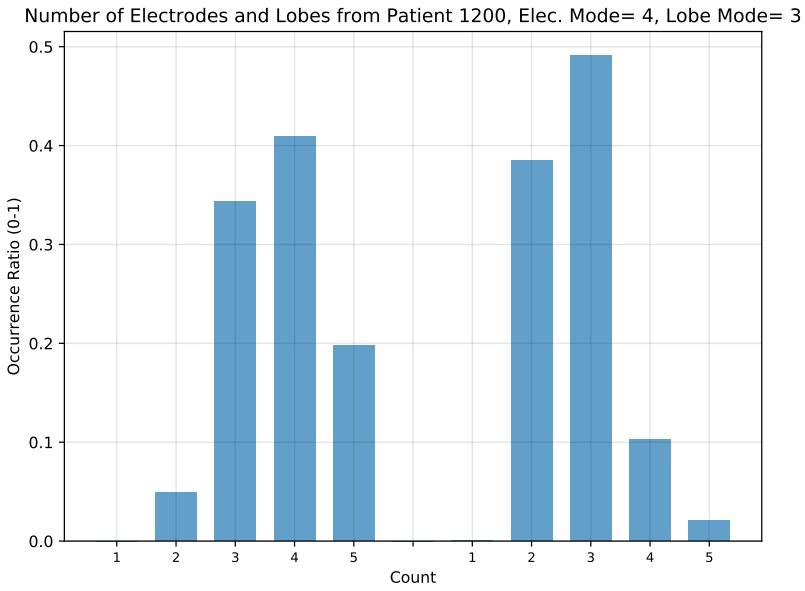


Figure S.1: The obtained electrode and lobe study, for all MOEA solutions, for patient 1200. On the left, the number of different electrodes. On the right, the number of different lobes.

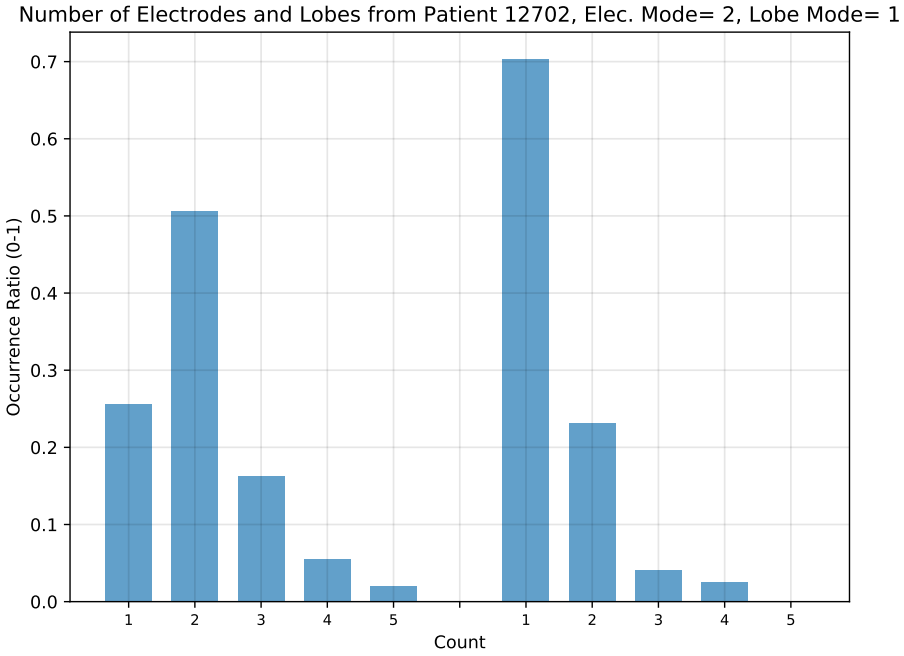


Figure S.2: The obtained electrode and lobe study, for all MOEA solutions, for patient 12702. On the left, the number of different electrodes. On the right, the number of different lobes.

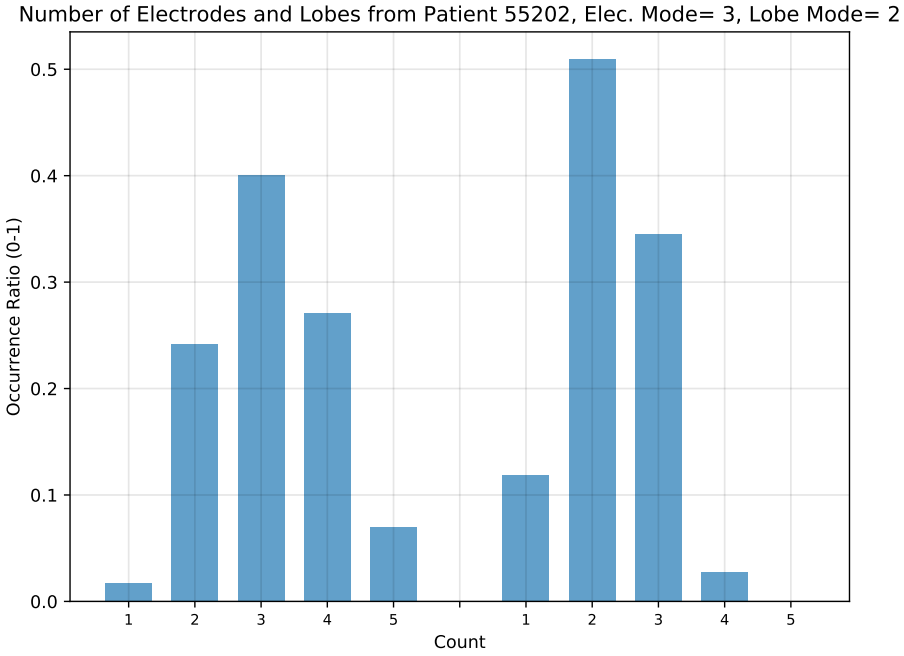


Figure S.3: The obtained electrode and lobe study, for all MOEA solutions, for patient 55202. On the left, the number of different electrodes. On the right, the number of different lobes.

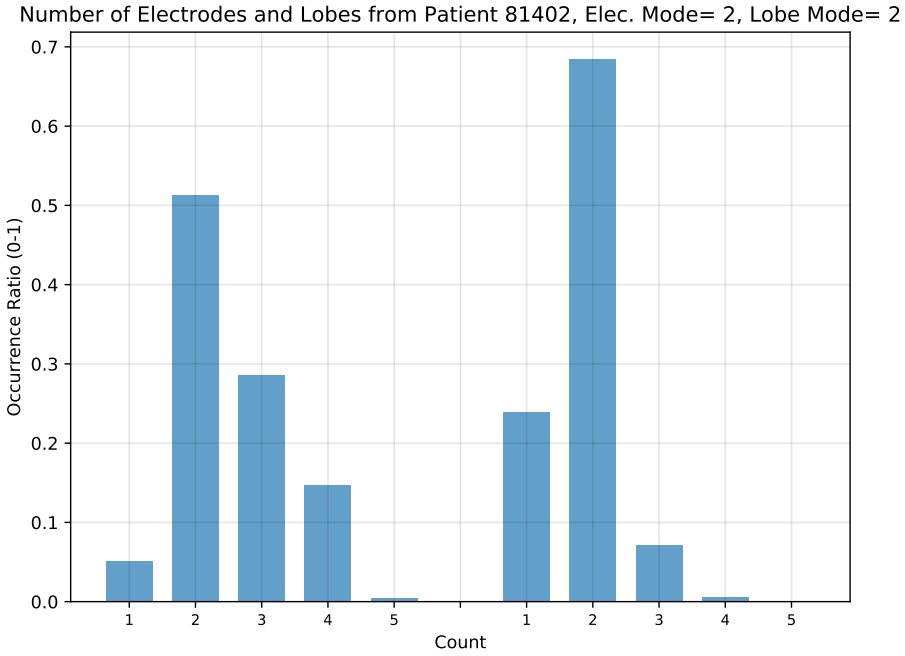


Figure S.4: The obtained electrode and lobe study, for all MOEA solutions, for patient 81402. On the left, the number of different electrodes. On the right, the number of different lobes.

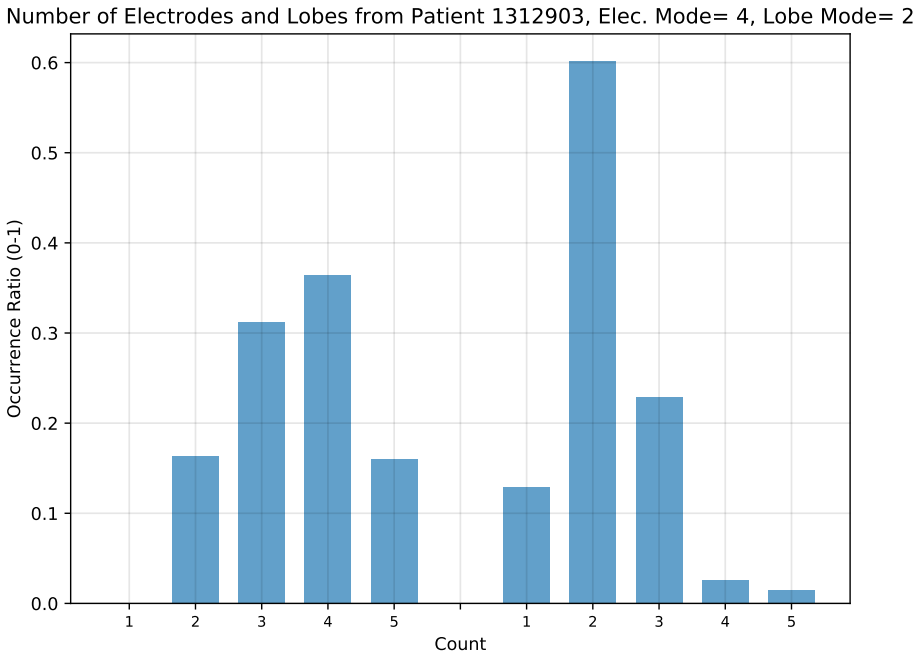


Figure S.5: The obtained electrode and lobe study, for all MOEA solutions, for patient 1312903. On the left, the number of different electrodes. On the right, the number of different lobes.

Appendix T

MOEA's impact of comfort in performance

The impact of patient comfort on performance, from chapter 5, is presented here. Three cases are possible: an increase (patient 1500), a maintenance (patient 21602), or a decrease (patient 58602).

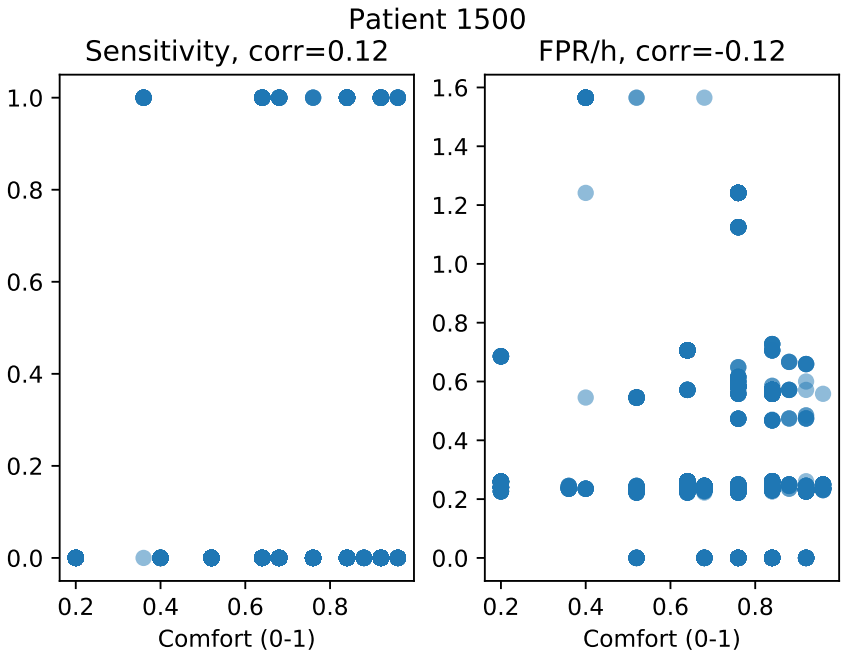


Figure T.1: The impact of comfort in performance for patient 1500. The overall performance increased.

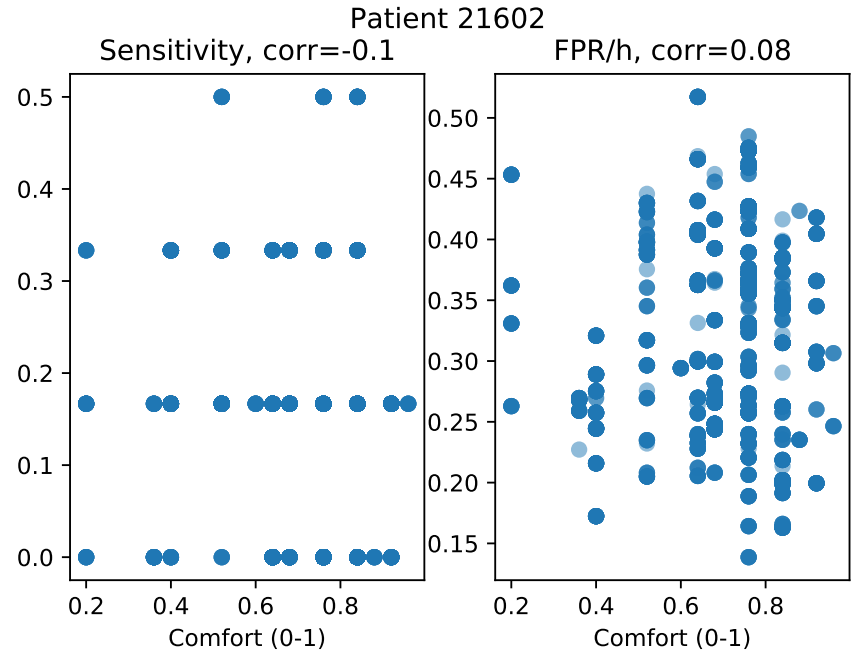


Figure T.2: The impact of comfort in performance for patient 21602. The overall performance was maintained.

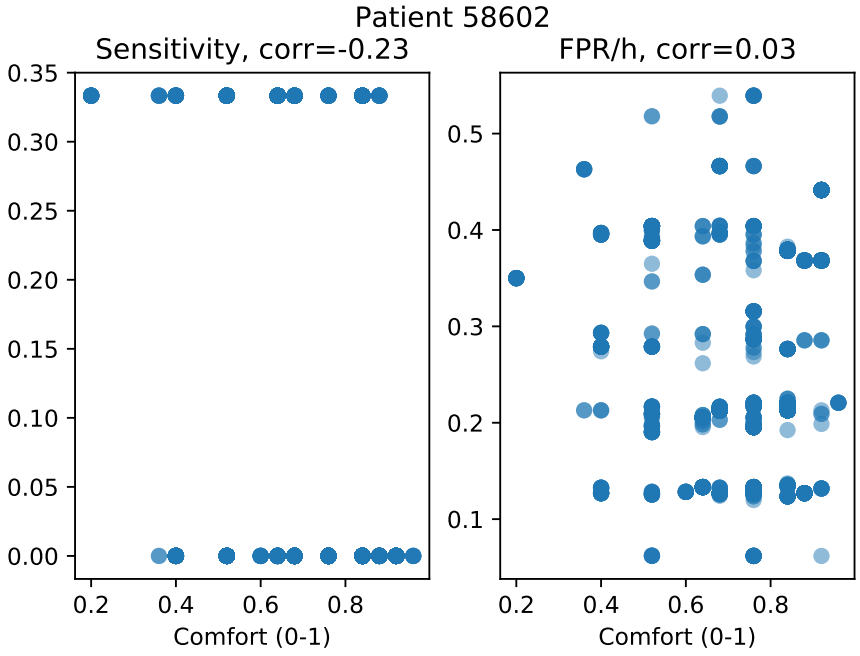


Figure T.3: The impact of comfort in performance for patient 58602. The overall performance decreased.

Appendix U

Patient information from the explainability study

The metadata about all analysed patients in chapter 6 is presented here.

Table U.1: Patients' information from the explainability study.

Patient ID	Sex AA	#Seiz. (Train/ Test)	Rec. duration (hours)	Lat.	Surgery decision	Seizure activity pattern	Seizure classification
402	f 55	3 2	98.14 23.07	l, r	i	t, t, t, t, t	FOIA, FBTC, FOIA FBTC, FOIA
8902	f 67	3 2	124.80 21.16	l	n	a, b, a m, a	UC, FOIA, FOIA FOIA, FOIA
11002	m 41	3 1	91.14 10.91	r	s	?, s, a t	UC, FOIA, FOIA FOIA
16202	f 46	3 4	96.82 31.93	l	n	r, ?, r, r r, ?, r	UC, FBTC, UC FOIA, FOIA, FOIA, FOIA
21902	m 47	3 1	62.56 9.42	l	s	t, t, t b	UC, FOIA, FOIA FOIA
23902	m 36	3 2	65.60 30.82	l	s	t, t, t d, t	FOA, FOA, FOA FOA, FOA
26102	m 65	3 1	58.13 21.88	l	i-n	m, t, t t	FOIA, FOIA, FOIA FOIA
30802	m 28	3 5	82.78 58.78	l, r	o	t, t, t t, t, t, t, t	FOA, FOA, FOA FOA, FOA, FOA, FOA, FOA
32702	f 62	3 2	109.94 18.78	l, r	n	t, t, t r, a	FOIA, FOIA, FOIA FOIA, FOIA
45402	f 41	3 1	68.24 18.54	l, r	i	t, t, t t	FOIA, FOIA, FOA FOIA
46702	f 15	3 2	46.20 11.89	r	o	a, a, t b, t	FOA, FOIA, FOIA FBTC, FOIA
50802	m 43	3 2	158.59 33.76	l	i-n	?, a, a t, t, t, t, t	FOIA, UC, UC FOIA, FBTC
52302	f 61	3 1	70.76 6.17	l	i-s	?, ?, d t	UC, FOA, UC UC
53402	m 39	3 1	65.73 12.25	l, r	s	?, ?, ? t	FOA, FOA, FOA FOIA
55202	f 17	3 5	45.32 62.46	l, b	i-n	t, d, t t, t, t, r, r	FOIA, FOIA, FOA UC, UC, FOA, UC, FOIA
56402	m 47	3 1	111.56 19.48	l, r	n	t, ?, ? a	UC, UC, UC FBTC
58602	m 32	3 3	92.68 22.65	l	i-n	r, t, t r, r, t	FOIA, FOIA, FOIA FOIA, FOIA, FOIA
59102	m 47	3 2	61.90 77.60	r	s	?, t, t t, t	FOA, FOIA, FOIA FOIA, FOA
60002	m 55	3 3	103.27 61.07	l, r	s	d, c, t t, d, d	FOIA, FOIA, FOIA UC, FOIA, FOIA
64702	m 51	3 2	71.45 30.15	r	i-s	?, m, t t, t	FOA, FBTC, FBTC FBTC, FBTC
75202	m 13	3 4	96.59 49.52	r	s	t, t, t t, t, ?, t	FOIA, FOIA, UC FOIA, FOIA, FOIA, FOIA
80702	f 22	3 3	46.29 28.62	b	n	b, b, ? c, c, c	FOIA, FOIA, UC FOIA, FBTCB, FOIA

Table U.1 continued from previous page

85202	f	3	52.01	l	s	m, c, m	FOIA, FOIA, UC
	54	2	19.61			m, m	UC, UC
93402	m	3	92.73	l	n	t, t, t	FBTC, FOIA, FOIA
	67	2	50.80			t, t	UC, UC
93902	m	3	189.32	r	i-n	t, t, d	FOA, FOIA, FBTC
	50	3	18.49			d, d, d	FOIA, FOIA, UC
94402	f	3	112.22	r	o	?, d, b	FOA, UC, FOIA
	37	4	28.36			t, ?, b, ?	UC, FOA, UC, FOA
95202	f	3	54.24	l	s	b, b, b	FBTC, FOIA, FOIA
	50	4	84.29			m, b, b, t	FOIA, UC, FOIA, UC
96002	m	3	46.05	l, r	s	t, t, t	FOIA, FOIA, FOIA
	58	4	67.38			d, a, t, a	FOIA, UC, FOIA, FOIA
98102	m	3	105.01	l	i-n	?, ?, ?	FOA, UC, UC
	36	2	43.36			?, ?	UC, FBTC
98202	m	3	105.94	r	o	t, a, t	FOIA, FOIA, FOIA
	39	5	45.44			t, t, t, t, t	FBTC, FOIA, FOIA, FOIA, UC
101702	m	3	27.28	l, r	n	t, t, t	FOIA, FOIA, FOIA
	52	2	22.62			r, r	FOIA, FOIA
102202	m	3	54.90	l	n	b, ?, t	FOA, UC, FOIA
	17	4	48.42			?, t, t, t	UC, FOA, FOIA, UC
104602	f	3	82.30	l	o	t, a, t	FOIA, FBTC, FBTC
	17	2	14.47			t, d	FBTC, UC
109502	m	3	72.73	l, r	n	t, t, t	FOIA, FOIA, UC
	50	1	39.11			t	UC
110602	m	3	86.03	r	o	t, t, t	FOIA, FOIA, FOIA
	56	2	24.87			t, t	FOIA, FOA
112802	m	3	68.61	l	n	t, t, t	UC, FOIA, UC
	52	3	103.64			t, t, t	FOIA, FOIA, UC
113902	f	3	57.85	r	o	t, d, t	UC, FOIA, FOIA
	29	3	20.78			t, t, t	FOIA, UC, FOIA
114702	f	3	45.87	r	o	t, t, t	FOIA, FOIA, UC
	22	5	31.67			t, d, t, d, t	FOIA, FOIA, FOIA, FOIA, FOIA
114902	f	3	25.68	l, r	n	s, b, s	FOA, FOIA, FOIA
	16	4	48.73			t, r, a, t	FBTC, UC, FOIA, FOIA
123902	f	3	146.71	l, r	i-n	t, t, t	FBTC, FBTC, FOIA
	25	2	29.21			t, t	FOIA FOA

Appendix V

Machine Learning pipelines from the explainability study

More details on the used Machine Learning pipelines for the three methodologies of the chapter 6, including the Convolutional Neural Networks (CNN) architecture, are presented here.

A grid search was performed with training seizures during training to find the preictal period (from 30 to 60 minutes). For the logistic regression and the Support Vector Machiness (SVMs) pipeline, the search for the k number of selected features was included. For the logistic regression, the best k features were selected through the F-test, a filter method that calculates the ratio between variances values [Venkatesh and Anuradha, 2019]. For the SVMs, an embedded forest of trees was used for selecting the best set of features for the following reasons: i) it is computationally light when compared to other embedded methods, and ii) since it is stochastic, it adds another layer of complexity which it was desired to retrieve explanations from complex methodologies. Finally, in the SVMs pipeline, the C cost-value was also searched.

Table V.1 summarises the grid search components of all pipelines.

Table V.1: Grid search components for each pipeline.

Pipeline	**Preictal Period (30-60 minutes)**	**k features (3,5,7,10,15,20,30)**	**C value (2**c, where c:-10:2:10)**
Log. Reg	x	x	
SVMs	x	x	x
CNNs	x		

For each set of search parameters (preictal period, number of features, and C value) and each of the three training seizures, a fitness value was obtained by using the geometric mean of sample sensitivity and sample specificity: $\sqrt{SS_{sample} * SP_{sample}}$. The geometric mean of each seizure was computed as follows: training seizures #1 and #2 and validating #3, training seizures #2 and #3 and validating #1, and

training seizures #1 and #3 and validating #2. Then, the three geometric means were averaged to get the fitness value of a given set of search parameters. The set having the highest fitness value was then selected.

Figures V.1, V.2, and V.3 show the full details of all pipelines. For the logistic regression feature selection, ANOVA F-test, scikit-learn's $SelectKBest(f\_classif, k = n\_features)$ function was used. For the feature selection in the SVMs pipeline, scikit learn's function $RandomForestClassifier$ with max_depth=10, $random\_state$=42, $n\_estimators$=100) was used. There were also used $linear\_model.LogisticRegression$ and $svm.LinearSVC$ functions from scikit-learn to train the logistic regression and SVM models.

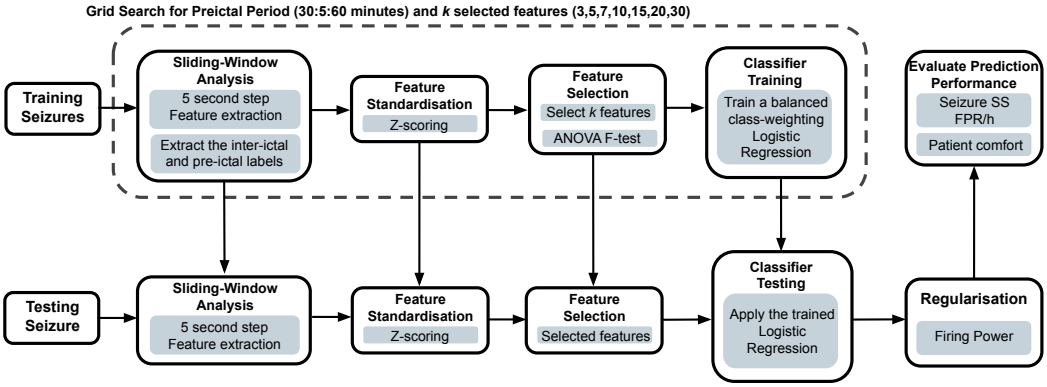


Figure V.1: Logistic regression model pipeline.

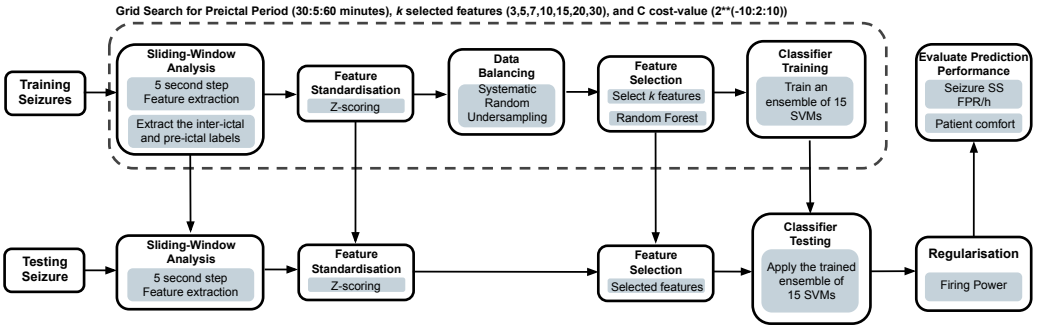


Figure V.2: Ensemble of the SVMs pipeline.

V.0.1 CNN architecture

This subsection provides the network architecture. A graphical version is depicted in Figure V.4.Three distinct convolutional parts were developed, each constituted by two convolutional 2D layers. The use of three convolutional components was a common architecture found in the literature [Usman et al., 2021a, Usman et al., 2020, Daoud and Bayoumi, 2019]. Filter size and values were found with an *a priori* grid search procedure. Convolutional layers with stride were used instead of max pooling layers as it would help the model learn automatically to reduce dimensionality instead of just performing a fixed operation.

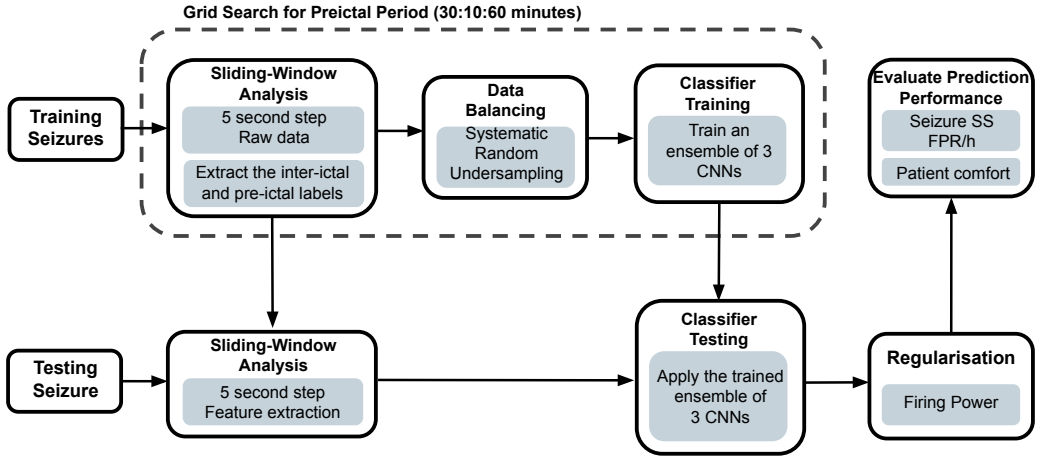


Figure V.3: Ensemble of the CNNs pipeline.

After each convolutional part, a 2D spatial dropout layer was applied along with an activation layer and batch normalisation. The dropout ratio was high (0.50) to avoid overfitting. Spatial dropout was used instead of regular dropout since pixels localisation is essential. In other words, neighbouring pixels correlate with each other as there is a spatial relationship. Therefore, spatial dropout drops entire feature maps instead of just one element, which helps the model to generalise. Batch normalisation layers are for stabilising training by re-centring and re-scaling. Swish [Ramachandran et al., 2017] activation function was used instead of Rectifier Linear Units (ReLU) to handle the dying neuron effect (when many ReLU neurons output a zero value, which may mainly happen to a learned negative bias).

A global average pooling 2D layer was used, followed by a dropout layer and a densely-connected layer. Lastly, it was used a softmax layer to convert the vector of values to a probability distribution.

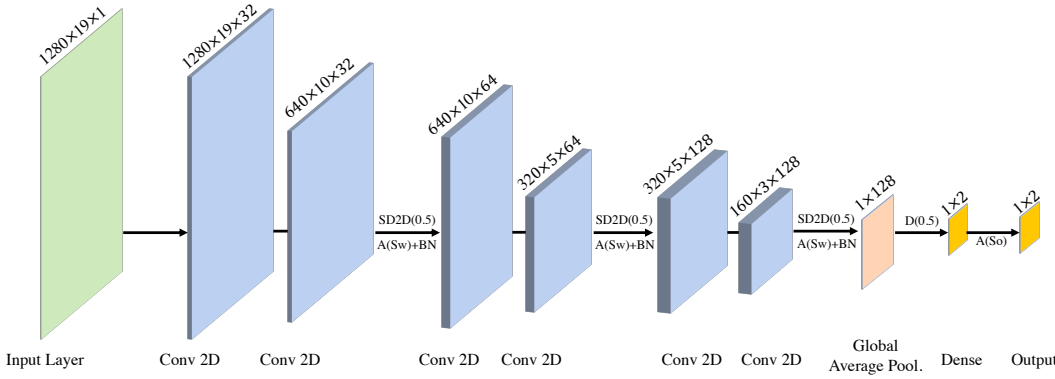


Figure V.4: CNN architecture. Conv2D stands for 2D Convolutional layer, SD2D(0.5) stands for 2D Spatial dropout layer with 0.5 dropout ratio, A(Sw) stands for Activation layer with swish function, BN stands for Batch Normalisation layer, D(0.5) stands for Dropout layer with 0.5 dropout ratio, A(So) stands for Activation layer with softmax function.

Appendix W

Seizure prediction results from the explainability study

The seizure prediction results for all patients, for all methodologies, from chapter 6 are presented here.

Table W.1: Prediction results for the logistic regression model. SS stands for Seizure Sensitivity and FPR/h for False Positive Rate per Hour. Performance above chance was analysed using surrogate analysis. Patients selected for explanations are written in bold.

Patient ID	Preictal Period (minutes)	#Features	SS	FPR/h	Above chance
402	**60**	**20**	**0.00**	**0.00**	**0**
8902	**35**	**7**	**1.00**	**0.11**	**1**
11002	30	7	0.00	0.78	0
16202	30	30	0.00	0.03	0
21902	50	10	0.00	0.00	0
23902	55	15	0.50	1.35	0
26102	60	30	0.00	0.00	0
30802	60	5	0.20	0.40	0
32702	30	7	0.50	0.06	1
45402	35	20	0.00	0.71	0
46702	40	20	0.00	0.00	0
50802	30	10	0.00	0.28	0
52302	55	30	0.00	1.01	0
53402	50	20	0.00	0.32	0
55202	55	3	0.20	0.55	0
56402	30	10	0.00	0.53	0
58602	30	3	0.00	0.54	0
59102	45	30	0.50	1.05	0
60002	45	7	0.00	0.13	0
64702	50	3	0.00	0.56	0

Table W.1 continued from previous page

75202	35	30	0.00	0.04	0
80702	50	30	0.33	0.28	1
85202	30	20	0.00	0.11	0
93402	**30**	**3**	**1.00**	**0.50**	**1**
93902	55	5	0.00	0.13	0
94402	55	30	0.00	0.80	0
95202	35	15	0.00	0.37	0
96002	35	3	0.25	0.69	0
98102	35	20	0.00	0.13	0
98202	35	30	0.00	0.02	0
101702	**60**	**30**	**0.00**	**0.71**	**0**
102202	60	3	0.00	0.05	0
104602	45	30	0.00	0.40	0
109502	40	20	0.00	2.32	0
110602	50	10	0.50	0.32	1
112802	30	3	0.33	0.75	0
113902	55	30	0.00	0.06	0
114702	45	30	0.00	0.00	0
114902	35	10	0.00	0.00	0
123902	35	3	0.00	0.00	0

Table W.2: Prediction results for the ensemble of SVM models. SS stands for Seizure Sensitivity and FPR/h for False Positive Rate per Hour. Performance above chance was analysed using surrogate analysis. Patients selected for explanations are written in bold.

Patient ID	Preictal Period (minutes)	#Features	SS	FPR/h	Above chance
402	60	30	0.00	0.00	0
8902	30	30	0.00	0.27	0
11002	30	10	0.00	3.93	0
16202	30	20	0.00	0.07	0
21902	45	15	0.00	0.28	0
23902	55	10	0.00	1.34	0
26102	60	30	0.00	0.10	0
30802	60	5	0.40	0.38	0
32702	30	20	0.00	0.12	0
45402	30	15	0.00	1.13	0
46702	**60**	**30**	**0.00**	**0.00**	**0**
50802	30	30	0.00	0.17	0
52302	60	3	0.00	1.13	0
53402	**60**	**3**	**1.00**	**0.22**	**1**

Table W.2 continued from previous page

55202	30	15	0.20	1.11	0
56402	30	15	0.00	3.83	0
58602	30	3	0.00	0.00	0
59102	**60**	**15**	**0.00**	**0.52**	**0**
60002	30	30	0.33	0.67	1
64702	50	3	0.00	1.02	0
75202	20	30	0.00	0.11	0
80702	45	30	0.33	0.44	1
85202	60	7	0.50	0.52	1
93402	30	7	0.50	0.77	0
93902	45	3	0.00	0.13	0
94402	30	15	0.00	1.29	0
95202	45	30	0.00	0.12	0
96002	55	20	0.00	1.52	0
98102	60	15	1.00	0.14	1
98202	30	30	0.00	4.14	0
101702	30	7	0.50	1.56	0
102202	60	5	0.00	0.39	0
104602	55	3	0.50	1.49	0
109502	60	7	0.00	0.65	0
110602	45	10	0.50	0.70	1
112802	45	3	0.00	3.01	0
113902	60	3	0.67	1.44	0
114702	45	5	0.00	0.04	0
114902	30	3	0.25	0.21	1
123902	30	3	0.00	0.00	0

Table W.3: Prediction results for the ensemble of CNN models. SS stands for Seizure Sensitivity and FPR/h for False Positive Rate per Hour. Performance above chance was analysed using surrogate analysis. Patients selected for explanations are written in bold.

Patient ID	Preictal Period (minutes)	SS	FPR/h	Above chance
402	60	0.00	0.00	0
8902	**60**	**0.50**	**0.00**	**1**
11002	60	0.00	0.00	0
16202	60	0.25	0.08	1
21902	60	0.00	0.00	0
23902	**30**	**0.00**	**1.65**	**0**
26102	30	0.00	0.10	0

Table W.3 continued from previous page				
30802	30	0.20	0.48	0
32702	**50**	**0.00**	**0.00**	**0**
45402	40	0.00	0.00	0
46702	30	0.00	0.10	0
50802	40	0.00	0.10	0
52302	60	0.00	0.00	0
53402	50	0.00	0.09	0
55202	60	0.00	0.07	0
56402	40	0.00	0.11	0
58602	40	0.00	0.05	0
59102	50	0.00	0.40	0
60002	50	0.00	0.20	0
64702	40	0.00	0.34	0
75202	30	0.00	0.16	0
80702	40	0.00	0.21	0
85202	30	0.50	0.37	1
93402	30	0.00	0.57	0
93902	45	0.00	0.13	0
94402	40	0.00	0.04	0
95202	30	0.00	0.08	0
96002	60	0.00	0.00	0
98102	40	0.00	0.00	0
98202	50	0.00	1.30	0
101702	30	0.00	0.10	0
102202	60	0.00	0.00	0
104602	40	0.00	0.08	0
109502	30	0.00	0.00	0
110602	50	0.00	0.00	0
112802	50	0.00	0.00	0
113902	30	0.00	0.11	0
114702	50	0.00	0.04	0
114902	30	0.00	0.07	0
123902	30	0.00	0.00	0

Appendix X

Interview script from the explainability study

Details about the interviews script concerning chapter 6 are presented here.

During the presentation to the interviewees, they were allowed to ask questions about any technical aspects or more details. These interviews were organic and informal. Thus, some questions were only asked when the interviewees had not spoken about some topics.

The following list shows some of the questions asked when the participants did not mention these topics beforehand. These topics were not discussed necessarily in this order.

- What do you think about the presented features?
- Do these features give you enough information?
- What do you think about the sequence of explanations? What do you think about their grouping and order?
- What do you think about the provided explanations? Which ones were more useful? Which ones were less useful?
- Are all these explanations too many?
- What do you think about the explanations about false alarms (patients 8902 and 93402)?
- What do you think about providing circadian and sleep-wake cycle information?
- What do you think about analysing the different 15 Support Vector Machiness (SVMs)' curves?

- Is there any issue or mistrust related to the models when we go from logistic regression for an ensemble of SVMs or Convolutional Neural Networkss (CNNs)?
- What are the limitations of the presented explanations?
- Would you like to have any other explanation?

In the end, the participants were always asked if they wanted to say anything else before finishing the interview.

Appendix Y

General patient analysis from the explainability study

The different typical cases, which were found during Firing Power and Concept Drifts (CDs) inspection in chapter 6 are presented here. Additionally, the statistical validation strategy for counting these methods is also presented here.

Y.1 Cases with good Firing Power but failed in prediction

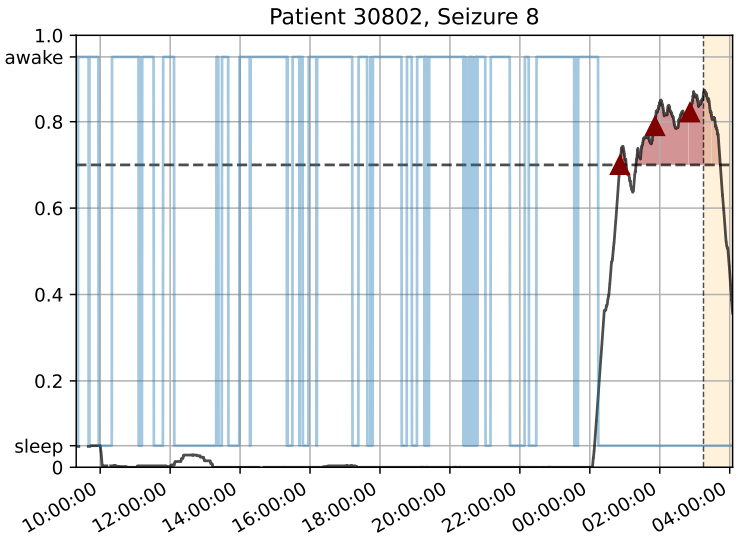


Figure Y.1: Seizure #8 of patient 30802. No alarm was raised during the preictal period due to the refractory period of each alarm. When analysing all available hours from that seizure, Firing power only started to rise when the patient went to sleep. Thus, all false alarms are close to each other and occur in the last four hours before the seizure onset.

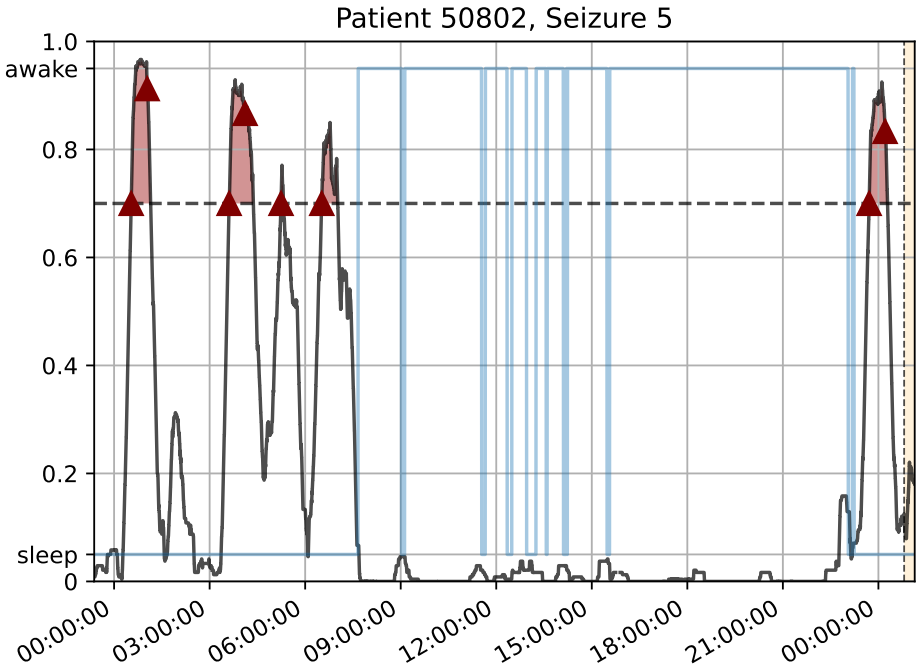


Figure Y.2: Seizure #5 of patient 50802. A clear firing power peak and two false alarms before the preictal period. Nevertheless, the classifier behaviour was good as these events occur relatively close to the seizure, and no other alarms occurred before in a near interval (no other alarms until about 17 hours before).

Y.2 Circadian-cycle influence in alarms and seizures

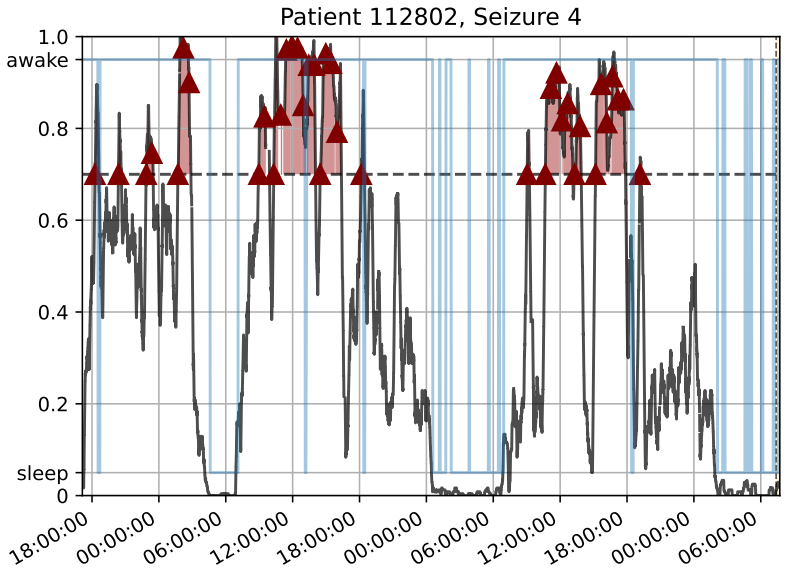


Figure Y.3: Seizure #4 of patient 112802. A false alarm cycle on two consecutive days: from 6am to 6pm. On the third day consecutive day, a seizure occurs after 6am.

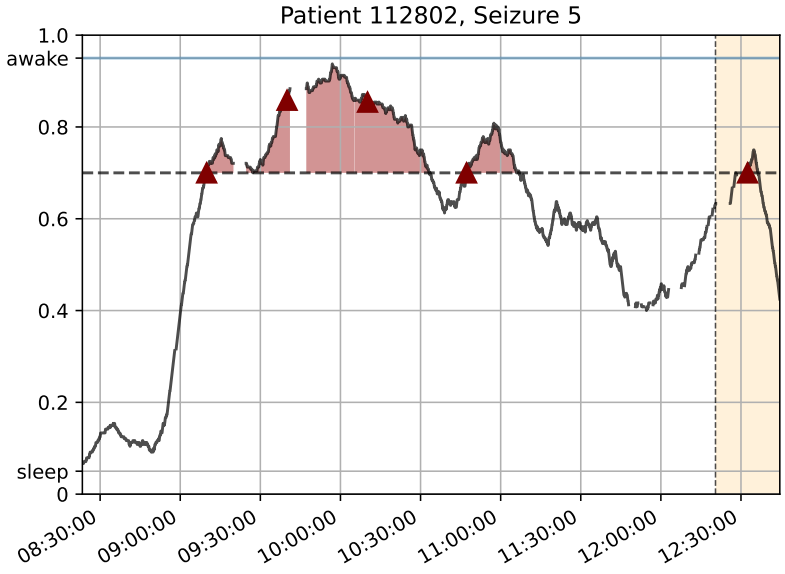


Figure Y.4: Seizure #5 of patient 112802. This seizure also occurs within the same cycle lasting from 6am to 6pm.

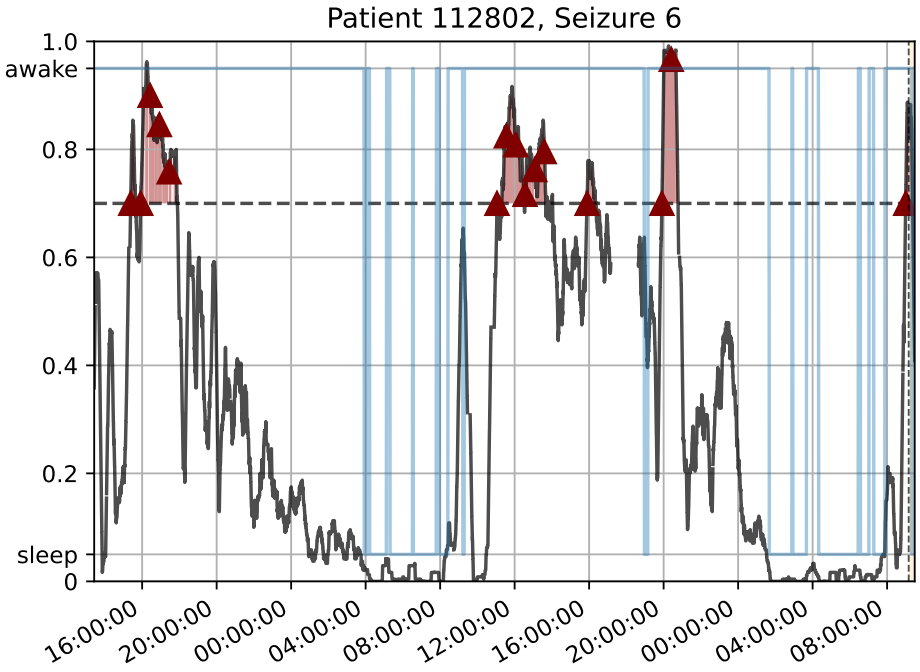


Figure Y.5: Seizure #6 of patient 112802. A shorter cycle is observed in this case: from midday to 4 pm. Then, the next day, a seizure occurs during the morning. All seizures and the majority of false alarms occurred during the mentioned cycle.

Y.3 Sleep-wake transition possible influence

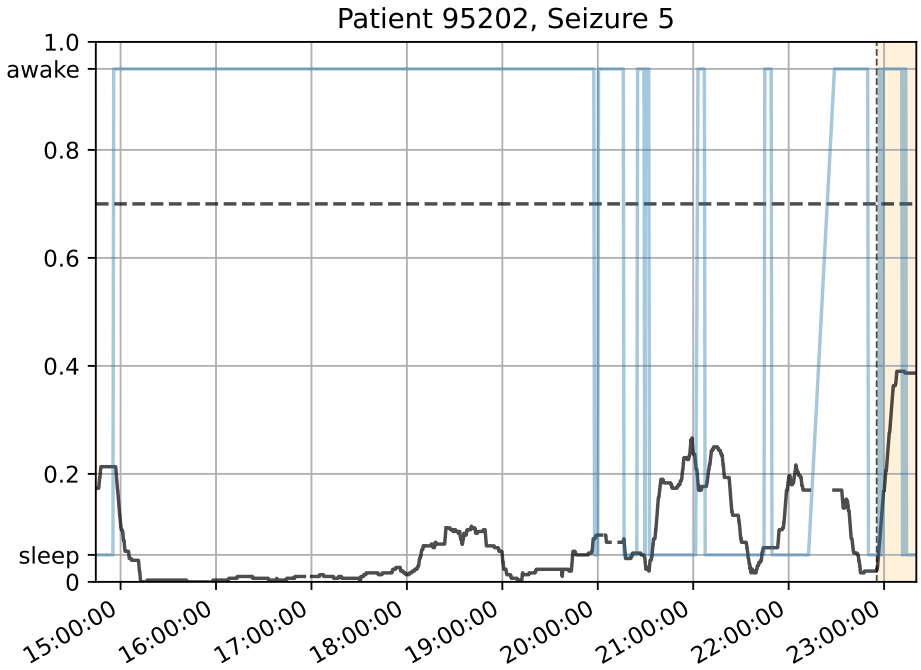


Figure Y.6: Seizure #5 of patient 95202. Visible influence of sleep-wake transitions. it is possible to see an influence on these transitions with small firing power peaks. When reaching almost midnight, a seizure occurred.

Y.4 Statistical validation

It was necessary to find a way to verify when a certain number of patients having a determined characteristic had statistical significance. In other words, when it was superior to a theoretical chance-level value.

A binomial distribution [Edwards, 1960] was used to model the number of successes in a sample of size n drawn with replacement from a population of size N. The binomial distribution may be a good approximation for an N much more extensive than n. The binomial distribution is the basis for the binomial test of statistical significance. The used of the binomial distribution was inspired in previous works [Pinto et al., 2022, Pinto et al., 2021a, Alvarado-Rojas et al., 2014], including the Multiobjective Evolutionary Algorithm (MOEA) chapter from this thesis.

Thus, to understand if the number of patients i) whose Electroencephalogram (EEG) forecasting was better/equal/worsen than the circadian forecasting or ii) whose circadian cycle influenced alarms and seizures, the cumulative binomial distribution (binomcdf) was used to verify the maximum number of successes given by chance-level. As the probability of success of these events was unknown, a p=0.05 was reasonably assumed. By looking for a significance α=0.05 and assuming a p=0.05, a number of successes whose probability was under 0.025 was looked for due to a Bonferroni correction [Weisstein, 2004] (inspired from multiple comparison testing). Although multiple comparison tests were not performed, this correction was still applied due to the p value.

By inspecting Figure Y.7, it was considered that four successes out of 40 were the maximum regarding chance level. More than four successes were considered to be above chance.

Concerning patients with at least one seizure presenting one determined characteristic (good firing power but did not predict the seizure or predictive sleep-wake transitions when the EEG model failed), some adaptations were made to this method. For each patient, there was the need to account for the possibility of occurring a particular characteristic in at least one seizure due to luck (by chance). Thus, for each patient, a p=0.05 was assumed and the probability of having success in at least one seizure by chance was calculated. The n was the number of tested seizures per patient. An average interpatient probability of 0.123 was obtained.

In Figure Y.8, an example was provided for a patient with three tested seizures, where a probability of success (by chance) is 0.142.

By using a similar rationale to the one in Figure Y.7, including an additional factor in the Bonferroni correction (probability should be lower than 0.0125), with n=40, x=0:1:40, and p=0.123, Figure Y.9 was obtained. By inspecting this figure, nine successes out of 40 are the maximum regarding chance level. More than nine successes were above chance.

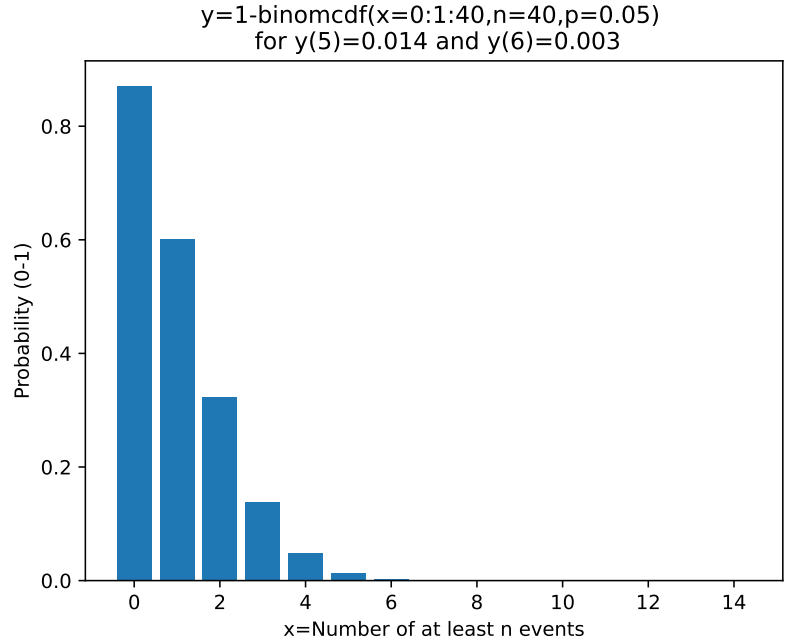


Figure Y.7: Graph used to access statistical significance where y=1-binomcdf(x,n,p). Y is the obtained probability (0-1), binomcdf is the cumulative binomial distribution, n is the number of patients (40), p is the probability of success, and x is a vector from 0 to 40 with step 1.

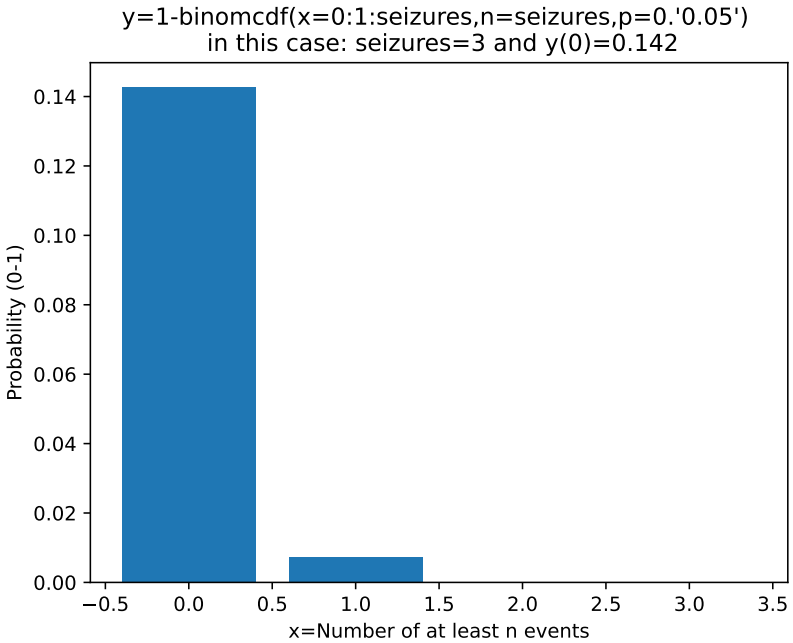


Figure Y.8: Graph used to access statistical significance where y=1-binomcdf(x,n,p). Y is the obtained probability (0-1), binomcdf is the cumulative binomial distribution, n is the number of seizures in this patient, p is the probability of success (0.05), and x is a vector from 0 to the number of seizures (3, in this case) with step 1. The obtained probability is y(0)=0.142.

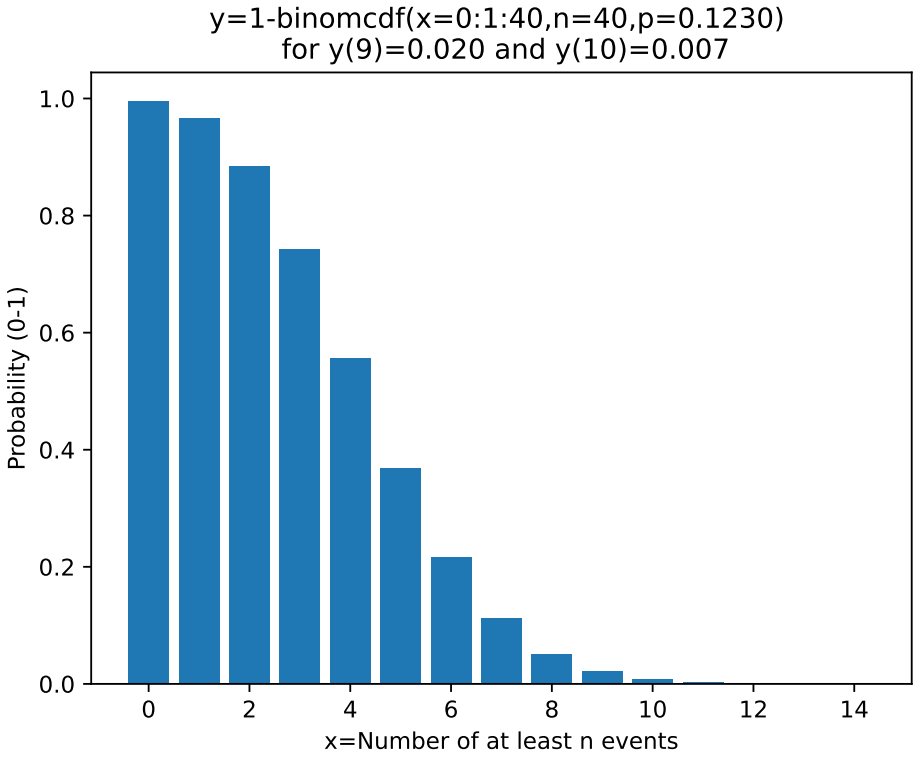


Figure Y.9: Graph used to access statistical significance where y=1-binomcdf(x,n,p). Y is the obtained probability (0-1), binomcdf is the cumulative binomial distribution, n is the number of patients (40), p is the probability of success(0.123), and x is a vector from 0 to the number of patients (40) with step 1.

Appendix Z

Models' comparison with a circadian forecasting algorithm

An illustration of the circadian forecasting algorithm and complete results from the forecasting algorithms, from chapter 6, are presented here.

The circadian forecasting algorithm only used circadian information: for each tested seizure, the circadian algorithm raised high seizure-risk warnings from 30 minutes before to 30 minutes after each seizure training onset time.

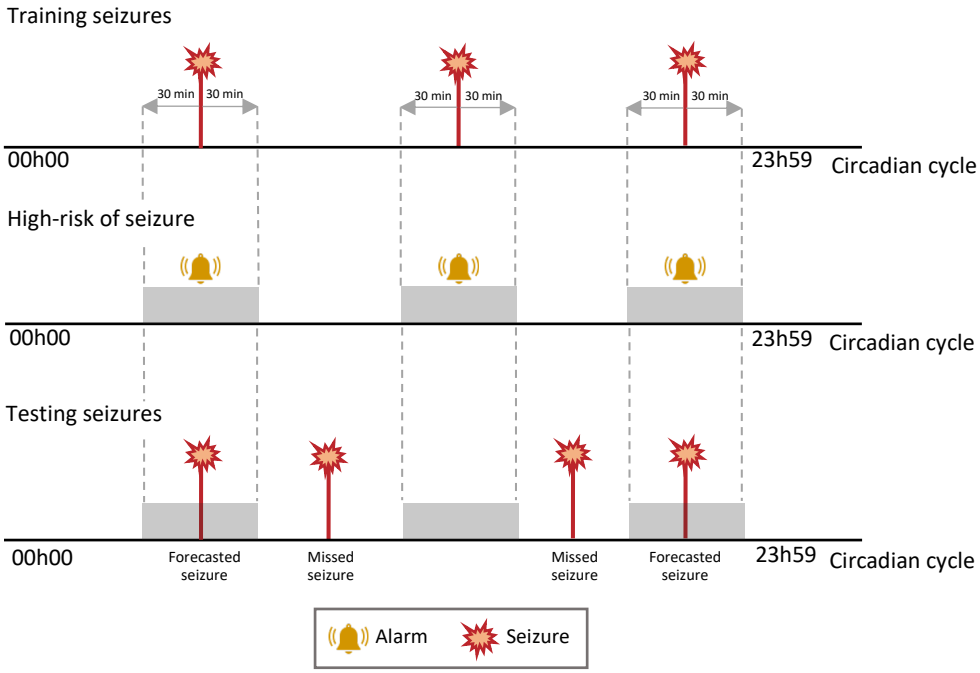


Figure Z.1: An example of the circadian forecasting algorithm.

A forecasting rationale also was applied to the logistic regression models, where high seizure-risk warnings corresponded to Firing Power values over the alarm

Table Z.1: The comparison of the EEG-based forecasting algorithm with a circadian forecasting algorithm.

	Seizure sensitivity					Time under warning (TiW)		
Patient	**Circadian model**	**EEG model**	**Circadian > EEG**	**Circadian = EEG**	**EEG > Circadian**	**Circadian model**	**EEG model**	**EEG time <Circadian**
402	0.00	0.00	0	1	0	3h00	0h00	1
8902	0.50	1.00	0	0	1	2h47	1h16	1
11002	0.00	0.00	0	1	0	3h00	1h46	1
16202	0.00	0.00	0	1	0	2h31	0h19	1
21902	0.00	0.00	0	1	0	3h00	0h00	1
23902	0.00	1.00	0	0	1	2h59	11h22	0
26102	0.00	0.00	0	1	0	2h40	0h00	1
30802	0.00	0.80	0	0	1	2h18	14h45	0
32702	0.00	0.50	0	0	1	2h09	0h25	1
45402	1.00	0.00	1	0	0	3h00	1h45	1
46702	0.50	0.00	1	0	0	3h00	0h00	1
50802	0.00	0.00	0	1	0	3h00	2h31	1
52302	0.00	0.00	0	1	0	3h00	2h36	1
53402	0.00	1.00	0	0	1	3h00	1h38	1
55202	0.40	0.80	0	0	1	3h00	16h26	0
56402	0.00	0.00	0	1	0	2h56	2h10	1
58602	0.33	0.00	1	0	0	3h00	2h47	1
59102	0.00	1.00	0	0	1	3h00	24h08	0
60002	0.00	0.00	0	1	0	2h22	3h29	0
64702	0.00	0.50	0	0	1	3h00	6h41	0
75202	0.00	0.00	0	1	0	3h00	0h43	1
80702	0.33	0.33	0	1	0	3h00	3h24	0
85202	0.00	0.50	0	0	1	3h00	0h31	1
93402	0.00	1.00	0	0	1	3h00	8h06	0
93902	0.00	0.33	0	0	1	3h00	0h55	1
94402	0.00	0.25	0	0	1	3h00	5h54	0
95202	0.00	0.25	0	0	1	2h07	8h18	0
96002	0.00	0.25	0	0	1	2h36	14h43	0
98102	0.00	0.50	0	0	1	3h00	1h16	1
98202	0.00	0.00	0	1	0	3h00	0h14	1
101702	0.50	0.50	0	1	0	3h00	7h28	0
102202	0.50	0.00	1	0	0	3h00	1h09	1
104602	0.00	0.00	0	1	0	3h00	1h47	1
109502	1.00	0.00	1	0	0	3h00	16h38	0
110602	0.00	0.50	0	0	1	3h00	4h15	0
112802	0.00	0.66	0	0	1	3h00	21h35	0
113902	0.33	0.00	1	0	0	3h00	0h46	1
114702	0.40	0.00	1	0	0	3h00	0h00	1
114902	0.00	0.00	0	1	0	3h00	1h46	1
123902	0.00	0.00	0	1	0	2h14	0h00	1
	Average SS		**Relative frequency**			**Average TiW**		**Relative frequency**
	0.15	0.29	0.18	0.40	0.420	02h52	01h32	0.62

threshold. These were then compared to the circadian forecasting algorithm, that only used circadian information.