On the clinical acceptance of black-box systems for EEG seizure prediction

Mauro. F. Pinto[1,*], Adriana Leal[1], Fábio Lopes[1], José Pais[2], António Dourado[1], Francisco Sales[3], Pedro Martins[1], and César A. Teixeira[1]

[1]Univ Coimbra, CISUC, Department of Informatics Engineering, Coimbra, Portugal
[2]Hospital CUF Tejo, Lisbon, Portugal
[3]Refractory Epilepsy Reference Centre, Centro Hospitalar e Universitário de Coimbra, EPE, Coimbra, Portugal
[*]mauropinto@dei.uc.pt

1 The seizure prediction ecosystem

In here, we present the seizure prediction ecosystem (the obtained network) with full detail, 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. We will explain these relations throughout this section while deepening parts that require more detail. In the end, we provide guidelines to help authors design their research. There is also available an interactive version as Supplementary Material which allows a free exploration and may be more intuitive.

Real life and Pre-Surgical Monitoring

We begin with the real life of an epileptic patient (1). Years after diagnosed with Drug-Resistant Epilepsy (DRE), a patient is referred to an epilepsy centre to undergo pre-surgical monitoring (5). The latter evaluates brain electrical activity (4) to localize the epileptic focus. If easily localized, removing the epileptic region is a possible solution [1, 2]. 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, we require patient consent (16→3) and an ethical justification (3). In this case, there is a strong motivation. Please note that this is a simplification of the pre-surgical monitoring process. We provide a more detailed explanation in the next subsection.

Despite pre-surgical monitoring is not as frequent as desired, happening for less than 1% of DRE patients, most studies are performed using pre-surgical monitoring data. Therefore, this data may not represent real life (2→5): the patient is in a controlled environment [3, 4]; the patient body may take time to adapt to the acquisition material (as initial data may need to be discarded) [5]; clinicians suppress medication to increase seizure occurrence frequency; and the short period, tipically, a couple of weeks of clinic admission and signal recording [6, 7] may mask the influence (1- -5) of day-to-day confounding factors (6- -4), such as stress, circadian and ultradian rhythms.

Most databases comprise pre-surgical 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 we will discuss later. Briefly, prospective studies require a significantly higher patient complacency, involve longer time periods, demand additional resources, and include higher risks for the patient. Prospective data then becomes retrospective (14- -7) [3, 4].

Pre-surgical monitoring details

Pre-surgical monitoring has the goal of successfully localize and the delineate the extension of the epileptogenic zone, ideally followed by a surgery to remove it. Towards this, clinicians begin patient analysis with a multimodal approach: long-term 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 delinate the epileptic region, other signals can be acquired, such as magnetic source imaging (MSI), functional MRI, SPECT, 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 epileptognenc zone can then be localized and be resected, the patient will undergo surgery. Otherwise, antiepileptic drugs, ketogenic diet, or neurostimulation are the possible current solutions [8].

In the literature, one can find different studies using data acquired during pre-surgical monitoring collected uing both scalp EEG and intracranial EEG (iEEG) [1, 9]. Thus, when comparing EEG seizure prediction among different types of EEG, it is relevant to understand and consider the situation that lead to the iEEG acquisition.

One must not forget that a patient is referred to a level 3 or 4 epilepsy centre [10] to do pre-surgical 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 [2].

Brain Dynamics

Brain dynamics (4) play a fundamental role in predicting seizures. Ictogenesis is known for leading to a hyperexcitability state that increases brain synchronization (see Figure 1). Thus, the EEG (4.1.1) is the most used signal. It can be acquired using scalp or intracranial (iEEG) electrodes, each one addressing different assumptions on brain dynamics and therefore being more compatible with specific applications [1, 11].

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 [1]. 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 [11]. 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

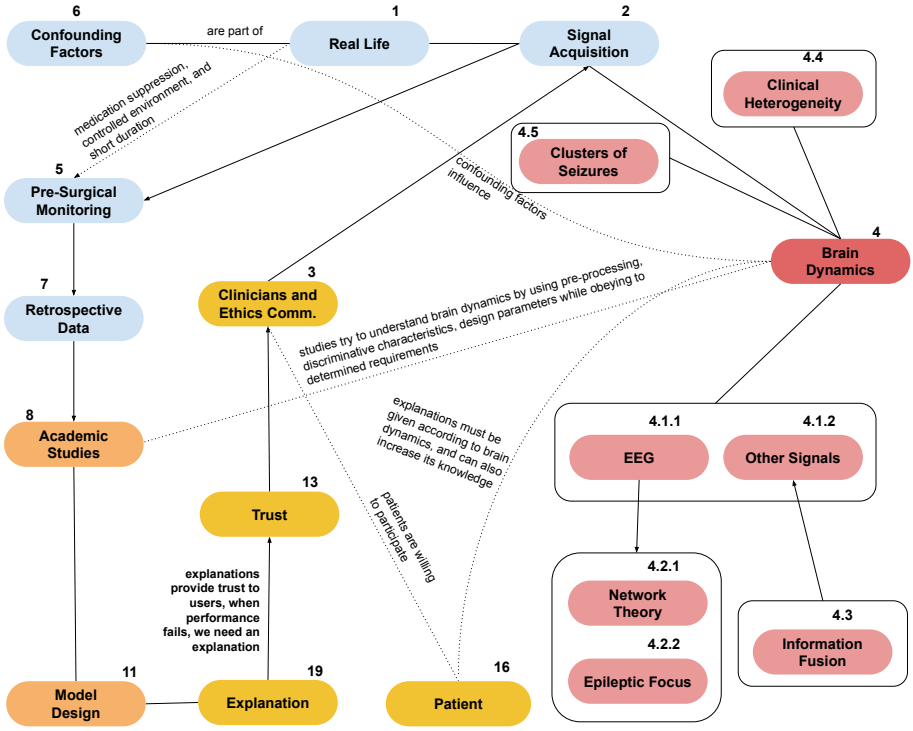

Figure 1: Details on the relations between actors concerning brain dynamics. Non-major actors are inside boxes.

or infection, among others [12]. Authors commonly focus on 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 [13] also explores EEG behind-the-ear that brings higher patient comfort, and Debener et al. [14] developed an EEG-ear array which demonstrated feasibility for long-term recordings.

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 pre-ictal interval [4]. For example, the cardiovascular dynamics regulated by the autonomous nervous system can be captured by the electrocardiogram and has been proven to carry complementary information for seizure prediction. Hence the growing belief that the analysis of multimodal data may provide improved results [4]. 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 directed at deepening understanding of this disease. There are several types of epilepsy syndromes, characterised by different types of epilepsy. 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 [15]. Seizure generating mechanisms are specific for each patient and each type of seizure [3, 4, 9], even though the source of spiking activity, for example, still remains unclear [1]. 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) [3, 4]. This aspect turns the ictogenesis process more complex and difficult to understand [**?**].

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 [4]. To understand academic studies, we also need to inspect Figure 2 for more details. Inevitably, we make several assumptions (see "Assumptions" section in this document for more information) when we design a new study. These may result from the used mathematical models, available data and other limitations, or even reflect the researcher knowledge concerning brain dynamics (8- -4).

Authors attempt to predict seizures by assuming the existence of the pre-ictal period. The latter is the transition between the normal brain state (inter-ictal period) and a seizure (ictal period). We can define the pre-ictal period in two different ways (8.1). One approach assumes it as a point of no return

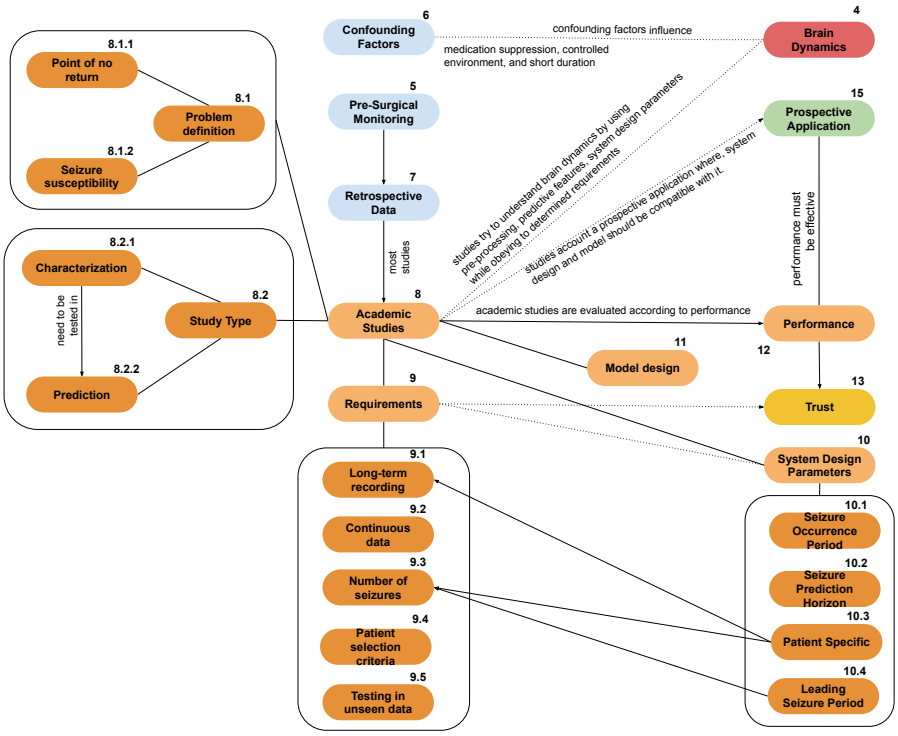

Figure 2: Details on the relations between actors concerning academic studies. System design parameters are also commonly named as "seizure prediction characteristics" in the literature [16]. Non-major actors are inside boxes.

(8.1.1), leading necessarily to a seizure [1]. 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 [5, 3]. These hypothesis influence significantly the experimental design, as it may be more difficult to have a ground truth or, in other words, a 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 context. The testing data requirements are: long term recordings (9.1), continuous data, without manually removing any segments due to noise or artifacts (9.2), 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) [1, 17, 4].

It is relevant to note the existence of two types of studies (8.2): characterisa-

tion (8.2.1) and prediction (8.2.2) [1]. In the first, authors try to find predictive models and/or predictive features that capture a clear distinct behaviour between a normal brain state and the pre-ictal period. However, the prediction potential of these should be further evaluated by integrating this information in 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 we focus here.

When considering a seizure intervention, system design parameters (10) have a significant role [1, 16]. 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 [3] 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 also limits the amount of data that can be used. Note that there is no definition/rule to consider a seizure as independent, which represents another difficulty regarding brain dynamics (4). Additionally, it is worth noting that, authors in prediction studies with pre-surgical monitoring data, tend to use shorter periods of time [9] for defining seizure independence comparing to a real life scenario [**?**].

Model Design

Figure 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 (inter-ictal or pre-ictal) throughout time. This model 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 [1, 17, 4, 9].

Before training a model, authors may pre-process (11.2) the signals to remove noise while maintaining the frequencies of interest and then, they extract predictive features (11.3) [17, 9]. 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 pre-processing (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 developing (18- -20). This

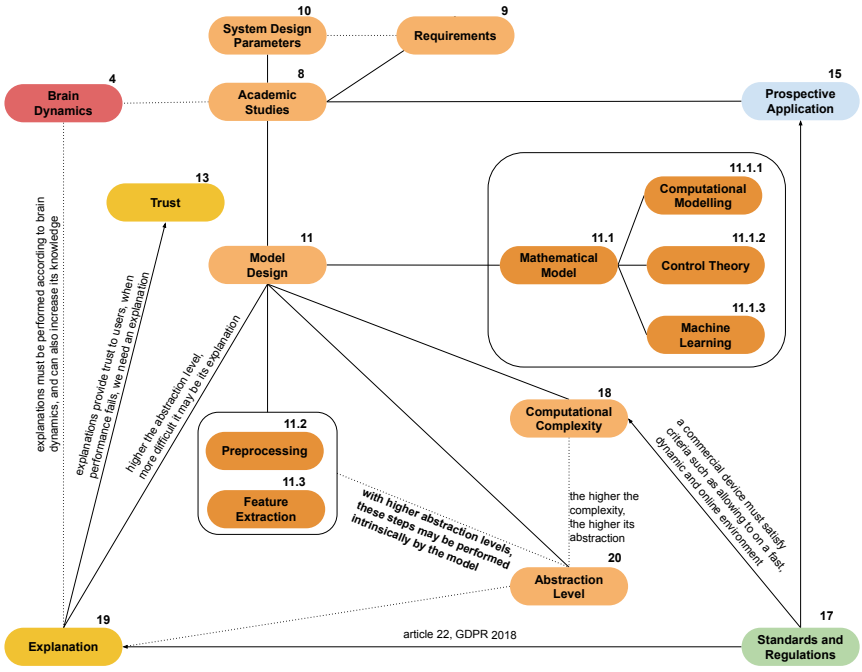

Figure 3: Details on the relations between actors concerning model design. Non-major actors are inside boxes.

can be an arising problem for real applications (17→18), as low computational requirements may be necessary [3, 4].

Although not mentioned directly, by choosing a given preprocessing method, feature, and model, we may be undertaking several assumptions on a physiological signal. Therefore, when constructing a pipeline, we challenge 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 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 (see in this document the "Assumptions" section).

Performance

Performance is one of the most discussed aspects in seizure prediction studies (see Figure 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 (FPR/h) [1, 17, 16]. Statisti-

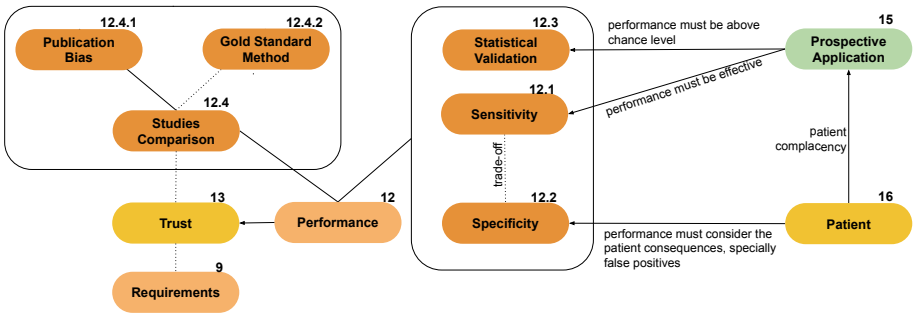

Figure 4: Details on the relations between actors concerning performance. Non-major actors are inside boxes.

cal validation (12.3) [4, 18, 19] 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 is the result of the identification of random phenomena in the biosignals rather than seizure-related patterns. This aspect becomes more relevant considering that seizure predition is a rare-event problem with considerable imbalance between inter-ictal and pre-ictal 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) [4, 16]. However, the results can be interpreted according to the envisioned clinical application, specifically by considering intervention consequences for patients (16→12.2). For instance, when considering the use of a warning system during pre-surgical monitoring, a maximum value of 0.15 FPR/h [1, 16] has been considered as the upper limit of false alarms that cause bearable/tolerable levels of stress and anxiety.

Studies comparison (12.4) enables to find methodologies that perform acceptably 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 [4, 9].

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 [1]. Nevertheless, it would be appropriate to compare results with a gold standard methodology applied in the same conditions [4].

Trust and Explainability

After proper studies comparison, one can ask what a good performance is, or even inquire about the minimum performance that justifies the design of a clin-

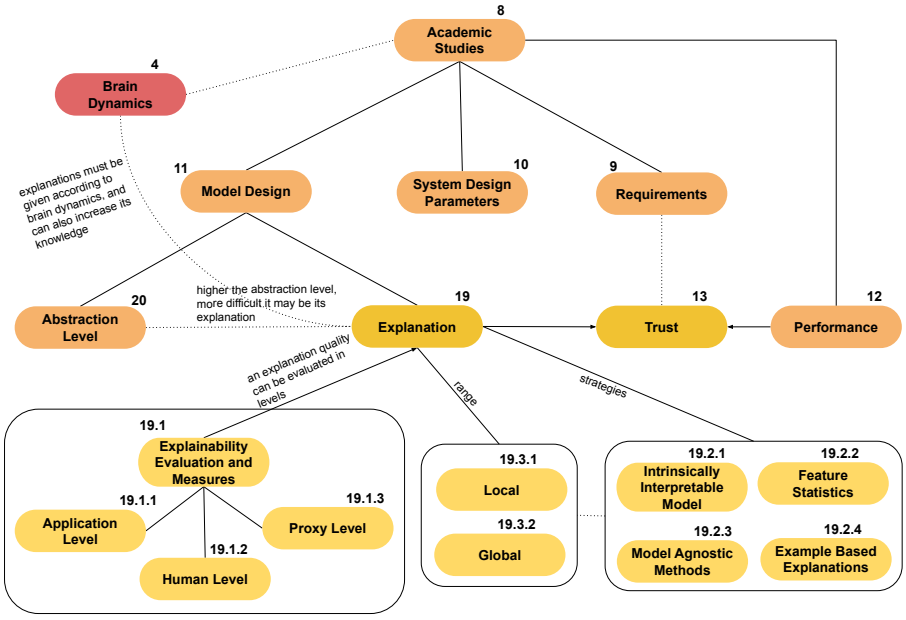

Figure 5: Details on the relations between actors concerning trust and explainability. Non-major actors are inside boxes.

ical trial. We believe that a proper methodology is the one which we 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 higher number of patient, trust increases as the testing data is more likely to represent real life conditions [4].

Although a given methodology, eventually, makes incorrect decisions, we can still trust it if one 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 [20]. 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 [3].

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 towards providing explanations (19- -20). Current clinical knowledge on 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 [21], its acceptance may differ among patients, clinicians, and data scientists. To better understand trust and explainability, we need to inspect Figure 5.

Explainability evaluation (19.1) is required. We can 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 in application and human levels [20, 22].

There are several strategies [20, 23] 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 visualization; iii) agnostic methods (19.2.3), which work on top of developed models [24, 25, 26, 27, 28, 29, 30]; and iv) example-based (19.2.4) by representing determined samples and showing the model decision [31, 32, 33, 34, 35]. The explanation range is also a topic of concern. It is local (19.3.1) when only explains a given decision for a sample and respective neighbourhood [20]. If it explains all samples, it is global (19.3.2).

Note that we did not consider a possible relation between patient and trust (16→13), as it concerns solely the algorithm design. Additionally, we also did not mention any connection between patient and explanation (16→19) directly, despite considering that a patient has the right for an adequate explanation concerning the device decisions. In fact, such rights are covered on article 22 from 2018 General Data Protection Regulation (GDPR) [36, 22, 31]. We believe that explanation and trust concern field experts, such as data scientists and clinicians. Nevertheless, patient comfort, trust and a proper explanation are fundamental. Therefore, we implicitly included these on the relation from 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 these 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. We can inspect details concerning prospective applications with Figure 6.

Ideally, studies using retrospective data envision and open the way to the enrollment in potential prospective scenarios (8- -15) [1, 11]. It is also possible to undergo a clinical trial without using any seizure intervention, as it happens with the ongoing SeizeIT2 clinical trial (NCT04284072). 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.

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) [37], responsive cortical stimulation as with the RNS®system

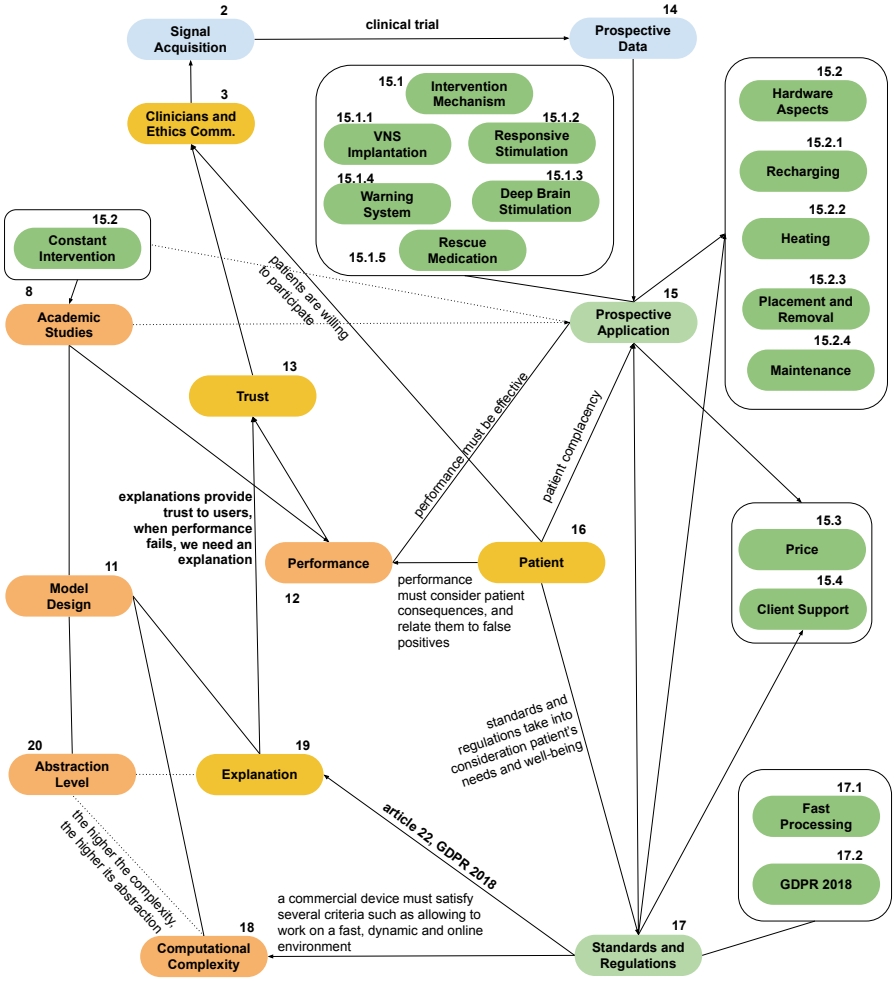

Figure 6: Detail on the relations between actors concerning a prospective application. Non-major actors are inside boxes.

(15.1.2) [12], or deep brain stimulation (15.1.3) [38]. The last was recently approved by the FDA [39] and encompasses two ongoing trials (NCT03900468, NCT02076698). An alternative could be a warning system (15.1.4) designed to minimise seizure consequences [1, 11] and/or taking seizure rescue medication, as benzodiazepines (15.1.5) [40, 41, 42]. Selecting an adequate intervention strategy is a complex task and must account for patient complacency and consequences (16→15).

It is interesting to reflect on the ideal scenario [1]. 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,

long 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 just focus on increasing knowledge on brain dynamics (15.2→8) as there was no need to investigate another prospective application. Given the amount of today's limitations, this may be utopic. However, we find it relevant to stress the purpose of seizure prediction research.

Naturally, device manufacturers must obey to 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). Others that are 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) [4, 11]. It is important to mention that considerable advances have been made in these devices, which is the case of IBM's neuromorphic TrueNorth chip [**?**] that already allows for deployment of deep learning models.

In fact, the price may be fundamental to the industry. Electrostimulation by implanting iEEG electrodes is currently considered the most promising strategy, as both RNS®system and Neurovista's system used iEEG [1, 5, 3, 4]. However, these may demand higher human and monetary resources than pre-surgical scalp EEG monitoring, which is already inaccessible to a large part of DRE patients. In the 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 [2, 10]. Thus, by focusing immediate efforts on low-cost and accessible warning systems followed by rescue medication intake, we may reach considerably more DRE patients.

The 2018 GDPR (17.2), for European citizens and European economic space [36], is 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 towards the patient safety (16→17).

2 Assumptions

Assumptions are a crucial part of any study for any scientific field. We often need to make assumptions about the world. Depending on the question we are asking, we may use a different perspective and therefore, different assumptions. In established areas, as the seizure prediction field authors may consider several assumptions that are not stated directly or even addressed properly in the discussion section. These assumptions may subconsciously be considered as part of public domain, particularly among peers. For non-experienced researchers, this may be a critical aspect.

Although we need to make assumptions to solve a problem, we should periodically review them. Table S1 presents what we considered to be 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 post-ictal 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 intracranial EEG, electrostimulation is usually the envisioned intervention. Thus, as the RNS system performs discharges up to 5000 ms, SOP periods must be short. If an author uses scalp EEG instead, a warning system is the envisioned intervention. Thus, 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 pre-processing, feature extraction, and/or model training. For example, when simply using 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, they choose the features with the highest discriminative power. With a regularization method (such as the LASSO regression), the best group of features is chosen, instead of the individual best. With a regularization method, authors also account for the interaction between features, by choosing the group with highest discriminative power (these may not have a high individual discriminative power). Thus, biologically speaking, regularization methods assume the possibility of existing more complex interactions in the brain when compared to filtering methods.

Approach	Assumption
Data	
Using pre-surgical monitoring data	Pre-surgical 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	
Pre-ictal 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 synchronization are probable. It may not lead to a seizure.
Fixed pre-ictal period	All seizures are generated in an equal window of time.
Study types	
Characterization	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	
Pre-Processing	The acquired signals have artefacts and noise that can be removed with pre-processing.
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.

Table S 1: Major assumptions on seizure prediction studies. Others are also possible, especially the ones concerning mathematical operations.

3 Questions about the seizure prediction future

Explanations and trust

Explanations help to detect data bias while increasing 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 among data scientists and clinicians:

1. Which are the concerns on explainability when designing seizure prediction models to prospective testing? Are clinicians sceptic about how the models work? Or are they afraid to compromise patient's 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 on 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 RNS system (NCT00572195) and the phase I NeuroVista Seizure Advisory System (NCT01043406), use algorithms with features that are clinically intuitive [3].

These two clinical trials demonstrate that, despite all the literature efforts put in developing complex models and consequent increase in performance, it may be necessary 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 a satisfactory performance on the application and human levels, can they be used?

5. Do we need a human-comprehensible explanation at every moment the algorithm is being used in real-time? Or do we need it only at certain moments, as with raised alarms and incorrect decisions? This may handle the fact that we, data scientists, tend to trust on model decisions when they are correct and only tend to inspect errors. In fact, when we train a machine learning model, we do it by minimizing errors of misclassified samples.

6. Can counterfactual explanations be interesting? Counterfactual explanations are very human-friendly and used widely by humans in daily life, because they can answer to a "why" question. This question can be formulated [20] as: what is the smallest change to the features that would change the prediction from alarm to no-alarm?

7. The used features in these studies [12, 5] (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, spectral relative 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, as input in the decision models, features that are clinically intuitive. 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 automatically perform feature extraction? In a positive scenario, would authors' methods need to explain which features were extracted by the Deep Learning model, or could an explanation simply consist in showing the relevant data points for a given decision?

Patients and real-applications

In the only (to the best of our knowledge) survey [43] on 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 our knowledge. This study was mostly a fixed questionnaire with few open questions on these parameters (SOP, 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 among data scientists, clinicians, and patients:

8. Could we obtain a different patient point of view, with the same subjects if we undergo a different approach, such as open questions only followed by a grounded theory analysis?

9. Despite their preferences, do patients have 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 its chance of happening? Moreover, the latter may lead to even higher monetary and psychological costs.

10. Concerning scalp EEG, few patients are willing to use long term acquisition systems. Should researchers make efforts in other formats of EEG scalp aquisition, as the two-electrode system from SeizeIT2 [13] or the ear-EEG array [14]? Or should they focus in other signals, despite having a lower theoretical potential, as the electroencephalogram (ECG)? For instance, smartwatches are more comfortable and can record an one-channel ECG. Are these strong reasons to promote the enrolling in long-term clinical trials using these devices, instead? They certainly allow more comfort and mitigate stigma, but its 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 [4, 1, 3] (10 minutes of SOP, minimum sensitivity of 90%, and very low FPR/h, simultaneously [43]). 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, we are excluding additional problems of device heating or energy consumption.

12. Should authors investigate the maximum false alarm rate that a patient can hold, without large physical and/or psychological consequences (due do too much electroelectroestimulation 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 pre-ictal activity at all? Efforts have already been made towards a proper system evaluation [44]. Would these electroestimulation systems benefit from using more robust algorithms to predict these sooner or are there seizures that brain electroestimulation can not disarm?

4 Social network iteration and refinement details

Figures 7-11 concern major iterations of the social network construction. Figure 11 concerns the complex network obtained before refinement and encapsulation. Please note that we also have changed some actors name and we renumbered them, during the refinement stage.

During network discussion, we also decided to add more details to some parts, as the explanation case. We added the evaluation levels, explanation range, and explanation strategies, found in Interpretable Machine Learning book [20] and in related articles [22]. Technological requirements and commercialization were also more detailed. We included i) hardware aspects, such as recharging, heating, placement and removal, maintenance, price, client support, fast processing, that can be found in Ramgopal et al. [11], and ii) information regarding GDPR's article 22 that can be found by analyzing Doshi Velez et al. [22] and Goodman et al. [36]. The GDPR is a clear case where we successfully inspected related articles within the initial ones [20], until reaching saturation. We also decided to highlight possible seizure interventions, found on several initial papers [11, 3, 4]. For the case of seizure interventions that deliver anti-epileptic drugs, we got input from the clinician that is also authoring this study regarding rescue medication such as diazepam. He advised us to search for epilepsy seizure rescue medication and also stressed the importance of epilepsy clinical heterogeneity, which we considered as well. Clusters of seizures (4.5) did not appear in the iteration models as it was included in the system requirements only. This was a codification limitation of our work which was successfully corrected by discussing the network among all authors of this study.

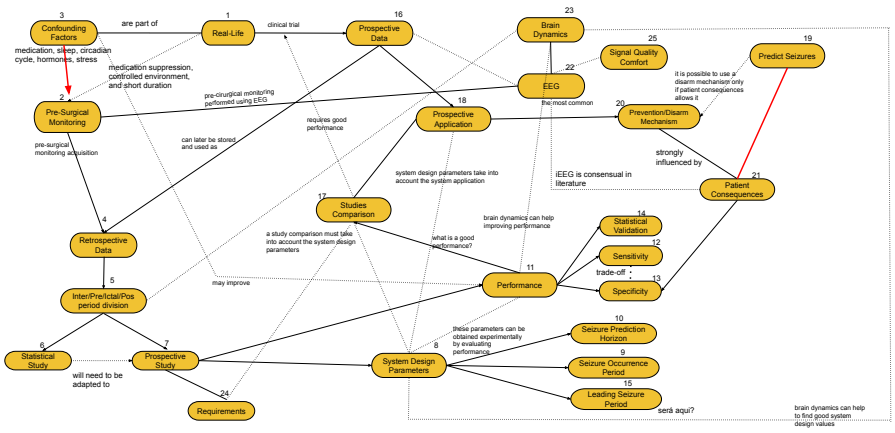

Figure 7: Social network iteration after analyzing Mormann et al. 2007 [1] and related articles. Red relations concern doubts raised at the time.

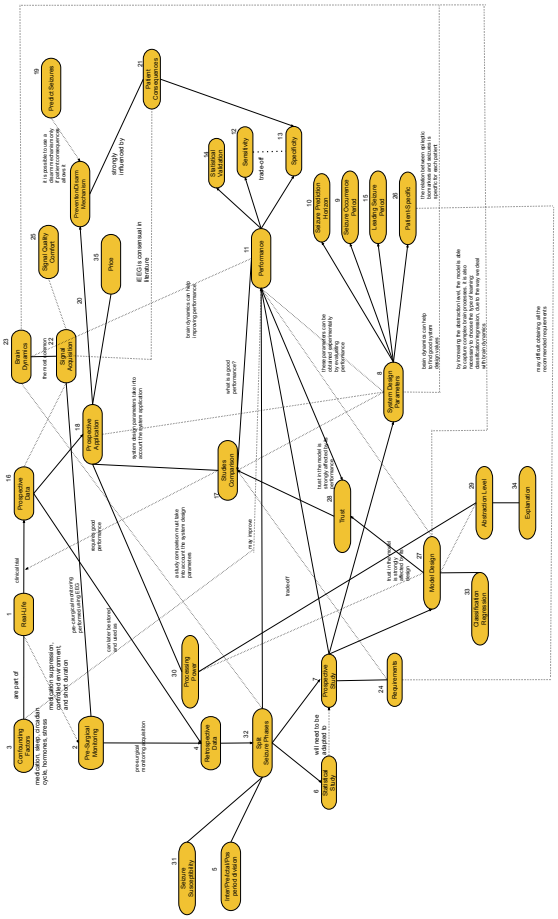

Figure 8: Social network iteration after analyzing Freestone et al. 2017 [3] and related articles.

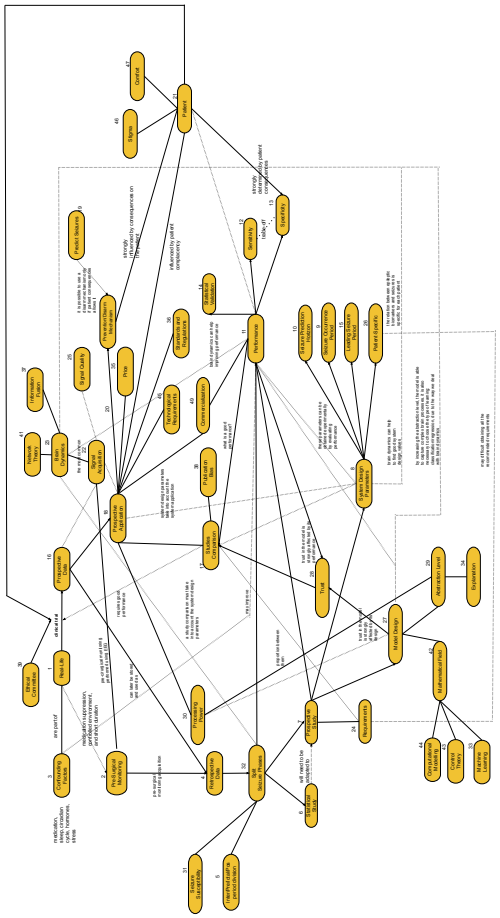

Figure 9: Social network iteration after analyzing Kuhlmann et al. 2018 [4] and related articles.

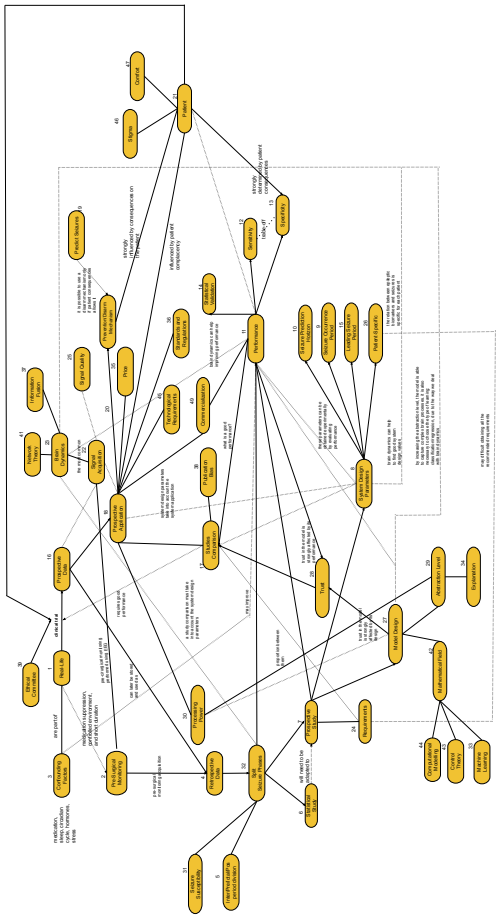

Figure 10: Social network iteration after analyzing Ramgopal et al. 2014 [11] and related articles.

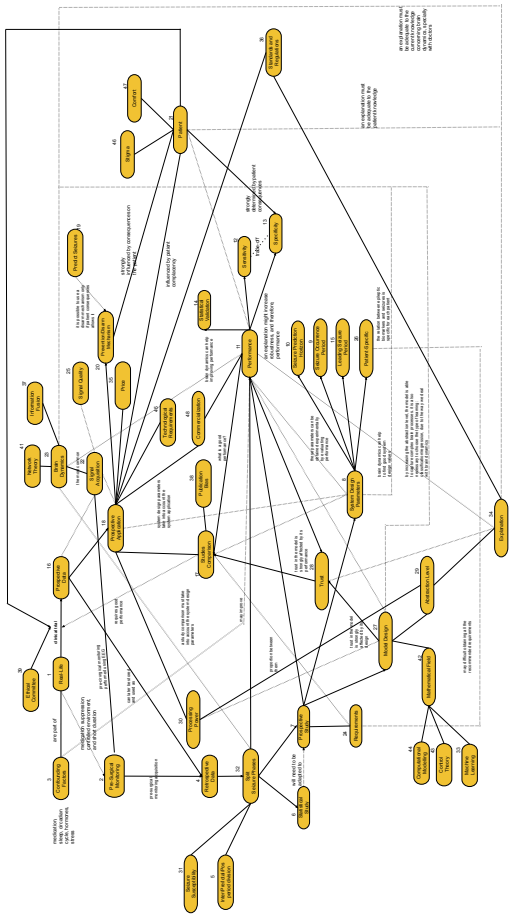

Figure 11: Social network iteration after analyzing Interpretable Machine Learning book [20] and related articles. Technical aspects on explainability evaluation and range are not present simply due to the figure size.

References

[1] Florian Mormann, Ralph G. Andrzejak, Christian E. Elger, and Klaus Lehnertz. Seizure Prediction: The Long and Winding Road. *Brain*, 130(2):314–333, 2007.

[2] Jerome Engel. What can we do for people with drug-resistant epilepsy?: the 2016 wartenberg lecture. *Neurology*, 87(23):2483–2489, 2016.

[3] Dean R Freestone, Philippa J Karoly, and Mark J Cook. A Forward-Looking Review of Seizure Prediction. *Current Opinion in Neurology*, 30(2):167–173, 2017.

[4] Levin Kuhlmann, Klaus Lehnertz, Mark P Richardson, Björn Schelter, and Hitten P Zaveri. Seizure Prediction — Ready for a New Era. *Nature Reviews Neurology*, 14(10):618–630, 2018.

[5] Mark J Cook, Terence J O'Brien, Samuel F Berkovic, Michael Murphy, Andrew Morokoff, Gavin Fabinyi, Wendyl D'Souza, Raju Yerra, John Archer, Lucas Litewka, et al. Prediction of Seizure Likelihood with a Long-Term, Implanted Seizure Advisory System in Patients with Drug-Resistant Epilepsy: a First-in-Man Study. *The Lancet Neurology*, 12(6):563–571, 2013.

[6] Philippa J Karoly, Hoameng Ung, David B Grayden, Levin Kuhlmann, Kent Leyde, Mark J Cook, and Dean R Freestone. The circadian profile of epilepsy improves seizure forecasting. *Brain*, 140(8):2169–2182, 2017.

[7] Maxime O Baud, Jonathan K Kleen, Emily A Mirro, Jason C Andrechak, David King-Stephens, Edward F Chang, and Vikram R Rao. Multi-day rhythms modulate seizure risk in epilepsy. *Nature communications*, 9(1):1–10, 2018.

[8] Chaturbhuj Rathore and Kurupath Radhakrishnan. Concept of epilepsy surgery and presurgical evaluation. *Epileptic disorders*, 17(1):19–31, 2015.

[9] Elie Bou Assi, Dang K. Nguyen, Sandy Rihana, and Mohamad Sawan. Towards Accurate Prediction of Epileptic Seizures: A Review. *Biomedical Signal Processing and Control*, 34:144–157, 2017.

[10] National Association of Epilepsy Centers. What is an Epilepsy Center? from www.naec-epilepsy.org. National Association of Epilepsy Centers Website in https://https://www.naec-epilepsy.org, 2021.

[11] Sriram Ramgopal, Sigride Thome-Souza, Michele Jackson, Navah Ester Kadish, Iván Sánchez Fernández, Jacquelyn Klehm, William Bosl, Claus Reinsberger, Steven Schachter, and Tobias Loddenkemper. Seizure Detection, Seizure Prediction, and Closed-Loop Warning Systems in Epilepsy. *Epilepsy & Behavior*, 37:291–307, 2014.

[12] Felice T Sun and Martha J Morrell. The rns system: responsive cortical stimulation for the treatment of refractory partial epilepsy. *Expert review of medical devices*, 11(6):563–572, 2014.

[13] Thijs Becker, Kaat Vandecasteele, Christos Chatzichristos, Wim Van Paesschen, Dirk Valkenborg, Sabine Van Huffel, and Maarten De Vos. Classification with a deferral option and low-trust filtering for automated seizure detection. 2020.

[14] Stefan Debener, Reiner Emkes, Maarten De Vos, and Martin Bleichner. Unobtrusive ambulatory eeg using a smartphone and flexible printed electrodes around the ear. *Scientific reports*, 5(1):1–11, 2015.

[15] Ingrid E Scheffer, Samuel Berkovic, Giuseppe Capovilla, Mary B Connolly, Jacqueline French, Laura Guilhoto, Edouard Hirsch, Satish Jain, Gary W Mathern, Solomon L Moshé, et al. Ilae classification of the epilepsies: position paper of the ilae commission for classification and terminology. *Epilepsia*, 58(4):512–521, 2017.

[16] M. Winterhalder, T. Maiwald, H. U. Voss, R. Aschenbrenner-Scheibe, J. Timmer, and A. Schulze-Bonhage. The seizure prediction characteristics: A general framework to assess and compare seizure prediction methods. *Epilepsy and Behavior*, 4(3):318–325, 2003.

[17] Kais Gadhoumi, Jean Marc Lina, Florian Mormann, and Jean Gotman. Seizure Prediction for Therapeutic Devices: A Review. *Journal of Neuroscience Methods*, 260(029):270–282, 2016.

[18] Björn Schelter, Matthias Winterhalder, Thomas Maiwald, Armin Brandt, Ariane Schad, Andreas Schulze-Bonhage, and Jens Timmer. Testing Statistical Significance of Multivariate Time Series Analysis Techniques for Epileptic Seizure Prediction. *Chaos*, 16(1), 2006.

[19] Ralph G. Andrzejak, Florian Mormann, Thomas Kreuz, Christoph Rieke, Alexander Kraskov, Christian E. Elger, and Klaus Lehnertz. Testing the Null Hypothesis of the Nonexistence of a Preseizure State. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 67(1):4, 2003.

[20] Christoph Molnar. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/, 2019.

[21] Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

[22] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[23] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*, 2019.

[24] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[25] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[26] Daniel W Apley and Jingyu Zhu. 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, 2020.

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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, 2016.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[29] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International Conference on Machine Learning*, pages 9269–9278. PMLR, 2020.

[30] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[31] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[32] Thomas Bäck, Mike Preuss, André Deutz, Hao Wang, Carola Doerr, Michael Emmerich, and Heike Trautmann. *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, 2020.

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[34] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29:2280–2288, 2016.

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

[36] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

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

[38] Paul Boon, Kristl Vonck, Veerle De Herdt, Annelies Van Dycke, Maarten Goethals, Lut Goossens, Michel Van Zandijcke, Tim De Smedt, Isabelle Dewaele, Rik Achten, et al. Deep brain stimulation in patients with refractory temporal lobe epilepsy. *Epilepsia*, 48(8):1551–1560, 2007.

[39] Epilepsy Foundation. 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, 2018.

[40] Robert C Tasker. Emergency treatment of acute seizures and status epilepticus. *Archives of disease in childhood*, 79(1):78–83, 1998.

[41] Marina Gaínza-Lein, Robert Benjamin, Coral Stredny, Marlee McGurl, Kush Kapur, and Tobias Loddenkemper. Rescue medications in epilepsy patients: a family perspective. *Seizure*, 52:188–194, 2017.

[42] Mark Scheepers, Bruce Scheepers, Michael Clarke, Susan Comish, and Matthew Ibitoye. Is intranasal midazolam an effective rescue medication in adolescents and adults with severe epilepsy? *Seizure*, 9(6):417–421, 2000.

[43] Andreas Schulze-Bonhage, Francisco Sales, Kathrin Wagner, Rute Teotonio, Astrid Carius, Annette Schelle, and Matthias Ihle. Views of patients with epilepsy on seizure prediction devices. *Epilepsy & behavior*, 18(4):388–396, 2010.

[44] Nathaniel D Sisterson, Thomas A Wozny, Vasileios Kokkinos, Anto Bagic, Alexandra P Urban, and R Mark Richardson. A rational approach to understanding and evaluating responsive neurostimulation. *Neuroinformatics*, pages 1–11, 2020.