

UNIVERSIDADE FEDERAL DE SANTA CATARINA  
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE - CAMPUS ARARANGUÁ  
DEPARTAMENTO DE COMPUTAÇÃO  
ENGENHARIA DE COMPUTAÇÃO

Guilherme Silva Inácio

**Data Mining of Intracranial Interictal EEG Recordings of Epilepsy Patients with Focal  
Cortical Dysplasia**

ARARANGUÁ

2019

Guilherme Silva Inácio

**Data Mining of Intracranial Interictal EEG Recordings of Epilepsy Patients with Focal  
Cortical Dysplasia**

Trabalho Conclusão do Curso de Graduação em  
Engenharia de Computação da Universidade Federal de  
Santa Catarina como requisito para a obtenção do Título  
de Bacharel em Engenharia de Computação  
Orientador: Prof. Dr. Alexandre Leopoldo Gonçalves  
Coorientador: Dr. Radek Janča

Araranguá

2019

Ficha de identificação da obra elaborada pelo autor,  
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Inácio, Guilherme Silva

Data Mining of Intracranial Interictal EEG Recordings  
of Epilepsy Patients with Focal Cortical Dysplasia /  
Guilherme Silva Inácio ; orientador, Alexandre Leopoldo  
Gonçalves, coorientador, Radek Janca, 2019.

108 p.

Trabalho de Conclusão de Curso (graduação) -  
Universidade Federal de Santa Catarina, Campus Araranguá,  
Graduação em Engenharia de Computação, Araranguá, 2019.

Inclui referências.

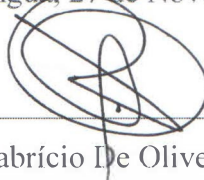
1. Engenharia de Computação. 2. Data Mining. 3.  
Epilepsy. 4. Electroencephalography. 5. Interictal  
Epileptiform Discharges. I. Gonçalves, Alexandre Leopoldo.  
II. Janca, Radek. III. Universidade Federal de Santa  
Catarina. Graduação em Engenharia de Computação. IV. Título.

Araranguá

**Data Mining of Intracranial Interictal EEG Recordings of Epilepsy Patients with Focal Cortical Dysplasia**

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de “Bacharel em Engenharia de Computação” e aprovado em sua forma final pelo Curso de Engenharia de Computação da Universidade Federal de Santa Catarina

Araranguá, 27 de Novembro de 2019.



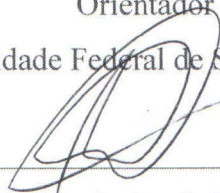
Prof. Fabrício De Oliveira Ourique, Dr.  
Coordenador do Curso

**Banca Examinadora:**

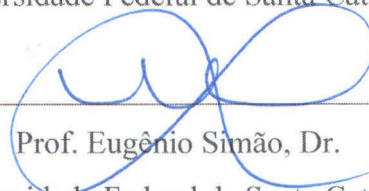


Prof. Alexandre Leopoldo Gonçalves, Dr.  
Orientador

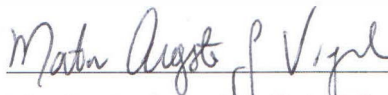
Universidade Federal de Santa Catarina



Prof. Antonio Carlos Sobieranski, Dr.  
Universidade Federal de Santa Catarina



Prof. Eugênio Simão, Dr.  
Universidade Federal de Santa Catarina



Prof. Martin Augusto Gagliotti Vigil, Dr.  
Universidade Federal de Santa Catarina

"Data are just summaries of thousands of stories – tell a few of those stories to help make the data meaningful."

(Chip & Dan Heath, 2014)

## ABSTRACT

Epilepsy is a group of neurological diseases that affects up to 1% of the world's population. About a third of the patients diagnosed with epilepsy are considered with difficult treatment (refractory), this group of patients can benefit from a resective surgery, that removes the epileptogenic tissue of the brain. Nowadays, the exams for the delineation of the areas for resection are still imprecise, and one of the techniques for a better definition of these brain areas require electrophysiological examination with invasive intracranial long-term electroencephalography monitoring (iEEG). One of the strategies for determining the epileptogenic zone (EZ) is to analyze the interictal data of patients with favorable outcomes and unfavorable outcomes with respect to the surgery resected areas and determine the statistical significance between them. A detection, analysis, and clustering data mining algorithm was used in order to extract information of 52 patients with focal cortical dysplasia (FCD) epilepsy. The detection algorithm identifies the interictal epileptiform discharges (IEDs) and arranges the detected activities into clusters given the patterns of spreading. For the statistical analysis, a comparison of the clustered data from three different vigilance epochs (sleep, awake and the combination of both) identified the most relevant epoch for identifying the epileptogenic areas and extract additional parameters. The results showed that the combined epoch of awake and sleep showed strong statistical significance in relation to the outcomes, followed by sleep and awake, respectively. Furthermore, given the positive results of the first analysis, an additional data mining was done in order to utilize the algorithm's outputs to predict the patient's FCD group, a predictive model was trained and displayed accuracy greater than 80% when tested with non-trained data.

**Keywords:** Data Mining. Epilepsy. Predictive analysis. Electroencephalography. Interictal Epileptiform Discharges.

## RESUMO

A epilepsia é um grupo de doenças neurológicas que afeta até 1% da população mundial. Cerca de um terço dos pacientes diagnosticados com epilepsia são considerados de difícil tratamento (refratários); esse grupo de pacientes pode se beneficiar de uma cirurgia ressectiva que remove o tecido epileptogênico do cérebro. Atualmente, os exames para o delineamento das áreas de ressecção em cirurgias ainda são imprecisos, e uma das técnicas para uma melhor definição dessas áreas cerebrais requer exames eletrofisiológicos com monitoração por eletroencefalografia intracraniana invasiva de longo prazo (iEEG). Uma das estratégias para determinar a zona epileptogênica (EZ) é analisar os dados interictais de pacientes com resultados cirúrgicos favoráveis e desfavoráveis em relação às áreas ressecadas da cirurgia e determinar a significância estatística entre eles. Um algoritmo de mineração de dados que realiza detecção, análise e agrupamento foi utilizado para extrair informações de 52 pacientes com epilepsia por displasia cortical focal (FCD). O algoritmo de detecção identifica as descargas epileptiformes interictais (IEDs) e organiza as atividades detectadas em grupos, dados os padrões de propagação. Para a análise estatística, uma comparação de dados de três estados de vigília (sono, vigília e a combinação de ambos) identificou o período mais relevante para identificar as áreas epileptogênicas e extrair parâmetros adicionais. Os resultados mostraram que o estado que combina vigília e sono mostrou forte significância estatística em relação aos resultados cirúrgicos, seguidos por sono, e depois vigília, respectivamente. Além disso, dados os resultados positivos da primeira análise, uma mineração de dados adicional foi feita para utilizar as saídas do algoritmo para prever o tipo de displasia cortical focal (FCD) do paciente. Um modelo preditivo foi treinado e exibiu precisão superior a 80% quando testado com um grupo de dados não treinados.

**Palavras-chave:** Mineração de dados. Epilepsia. Análise preditiva. Eletroencefalografia. Descargas epileptiformes interictais.

## RESUMO EXPANDIDO

### INTRODUÇÃO

Epilepsia é considerada um grupo de doenças neurológicas crônicas que afetam entre 0,5% e 1% da população mundial (BANERJEE; HAUSER, 2008). É caracterizada por convulsões recorrentes não provocadas, que podem ser descritas pelo comportamento atípico das funções cerebrais causadas por descargas elétricas anômalas no cérebro (FISHER et al., 2005). As convulsões variam de lapsos de consciência, movimentos corporais involuntários, distúrbios sensoriais até convulsões prolongadas e severas. A doença pode causar diversas consequências neurológicas, cognitivas, psicológicas e sociais na vida da pessoa, afetando negativamente sua qualidade de vida (FISHER et al., 2014). Tratamentos incluem dietas e fármacos, mas cerca de um terço dos pacientes são considerados fármaco-resistentes, também conhecidos como refratários. O tratamento para o subgrupo de pacientes refratários inclui procedimentos cirúrgicos no cérebro para a retirada do tecido epileptogênico (TANG et al., 2017). A epilepsia pode ser congênita ou adquirida, sendo a primeira causada por malformação cerebral durante a formação do feto, e a última sendo originada por vários fatores, incluindo lesões na cabeça, infecções e tumores cerebrais (SYNAPSE, 2019). As malformações do desenvolvimento do córtex cerebral são uma das principais causas de epilepsia do tipo refratário, sendo a Displasia Cortical Focal (FCD) uma delas. Para a FCD, as anomalias estruturais no desenvolvimento do córtex cerebral são classificadas em três tipos (I, II e III) e podem ser detectadas por técnicas de diagnóstico por imagem como a ressonância magnética e o eletroencefalograma (KABAT; KRÓL, 2012). Um dos problemas atuais é como definir com precisão as estruturas epileptogênicas cerebrais, sendo este um problema multidisciplinar envolvendo áreas da medicina e bioinformática. Partindo da área médica, temos a definição da zona epileptogênica (EZ) como: “a menor quantidade de tecido cortical que deve ser ressecado (inativado ou completamente desconectado) para garantir ausência de convulsões no pós-cirurgia.”, além das sub-estruturas que a caracterizam como: a zona de início da convulsão (SOZ), zona irritativa, zona sintomatogênica e a lesão epileptogênica. Técnicas de análise de sinais para a correta detecção da EZ estão sendo pesquisadas e desenvolvidas nos últimos dez anos. O entendimento atual da área, contrariando o antigo entendimento, é de que as atividades epiléticas não são focalizadas, mas sim o resultado de uma complexa interação de redes cerebrais (BARTOLOMEI et al., 2017). Com o intuito de modelar e entender essa complexa rede cerebral, diversos métodos estão sendo propostos, de análise de grafos (WILKE et al., 2011) ao uso de algoritmos de aprendizado de máquina (GRINENKO et al., 2018), com resultados variados. Recentemente um método de mapeamento da organização cerebral, descrevendo-a de uma forma modular, caracterizada por diversas sub-regiões (*clusters*), demonstrou que para 12 de 14 pacientes a sub-região de maior atividade cerebral correspondia a SOZ (JANCA et al., 2018). Considerando o algoritmo proposto por Janca et. al. (2018), e devido a fato que ele demonstrou bons resultados e deixa espaço para validações e trabalhos derivados, temos a seguinte pergunta de pesquisa: “Como extrair parâmetros representativos dos resultados do método escolhido, quando aplicado a dados de pacientes com FCD, de modo a ajudar no melhor delineamento da zona epileptogênica?”.



## OBJETIVOS

O objetivo geral deste trabalho é identificar e verificar parâmetros representativos da atividade epilética, extraídos de dados de eletroencefalogramas intra-craniais, de modo a ajudar a definir a estrutura epilêptogênica crítica no cérebro. Para isso, é preciso verificar a significância estatística do algoritmo escolhido para os dados selecionados, comparando-a com resultado da cirurgia, com o auxílio de um índice. Também é necessário verificar as diferenças entre as épocas de vigília para os dados obtidos, de modo a definir qual a época mais significativa para análise e calcular um limite que melhore a significância estatística do algoritmo. E por último, descobrir e testar a efetividade dos dados de saída para detectar o tipo de FCD do paciente, utilizando técnicas de aprendizado de máquina.

## METODOLOGIA

A metodologia desse trabalho dá-se da seguinte forma. Os dados coletados no Hospital Universitário da Universidade Charles foram disponibilizados para análise pelo grupo de pesquisa em análise de sinais da Universidade Tecnológica Tcheca em Praga. Eles são compostos de gravações de eletroencefalogramas intra-craniais (iEEG) em conjunto com dados médicos do paciente, como definição clínica da zona de início da convulsão e área ressecada do cérebro. Os dados brutos são inseridos em um algoritmo de mineração de dados de EEG composto de diversas etapas. Dentre elas, a detecção de descargas epileptiformes interictais (IED) usando modelos estatísticos de análise de sinais (JANCA et al., 2015), a posterior clusterização das IED detectadas de acordo com os padrões de início e propagação no cérebro (JANCA et al., 2018), e por fim a seleção manual dos *clusters* detectados. Com as saídas desse algoritmo, dá-se início ao processo de cálculos de índices e verificação estatística utilizando-se dos dados médicos de resultado da cirurgia. A análise é repetida para cada período de vigília, de forma a detectar o período que oferece a melhor correlação estatística entre o resultado da cirurgia e o índice calculado. Um limite inferior para os *clusters* selecionados é calculado de forma a melhorar a significância estatística obtida. Com o melhor cenário definido, faz-se uma nova análise, utilizando técnicas de aprendizado de máquina, objetivando verificar se é possível prever o tipo de displasia cortical focal do paciente, baseando-se nos dados extraídos do algoritmo.

## RESULTADOS E DISCUSSÃO

Para os três períodos de vigília analisados (sono, vigília e a combinação de ambos), tem-se que para todos se observa significância estatística entre o índice de ressecção e o resultado da cirurgia. O melhor resultado dentre todos os cenários analisados foi o período combinado de sono e vigília, com um limite inferior de 10% para os *clusters*, onde se obteve um *p*-valor de .00198, sendo este o período que proporciona maior informação de localização da zona epileptogênica. Os resultados alcançados vão ao encontro de outra pesquisa recente (PETR KLIMES, et al., 2019), com concordância parcial sobre a ordem de significância dos estados

de vigilância. Para a segunda parte da análise, utilizando-se dos dados do período mais significativo para o treinamento de um modelo *Ensemble* de classificação, obteve-se uma precisão da predição do tipo de FCD do paciente maior que 80% para o grupo de teste, e maior que 85% quando testado em pacientes com FCD do tipo III. O que demonstra que é possível extrair e detectar padrões significativos com os resultados do algoritmo de mineração de dados utilizado. No entanto, a amostra de pacientes para essa análise foi menor que a recomendada, o que diminui a validade dos resultados da precisão do modelo.

## CONCLUSÃO

Neste trabalho foram realizados dois experimentos, o primeiro apresentou uma análise estatística do eletroencefalograma de pacientes com displasia cortical focal, utilizando um algoritmo de mineração de dados que detecta, analisa e agrupa IEDs por seus padrões de dispersão. O objetivo da análise foi investigar a associação entre os dados analisados e os resultados cirúrgicos dos pacientes. O conjunto das saídas do algoritmo com as informações médicas permite o cálculo de um índice de ressecção e inspeção da correlação estatística entre eles. Além disso, examinando os diferentes períodos de vigilância, foi possível identificar os mais relevantes para a análise, o que poderia oferecer informações mais pertinentes para um melhor delineamento da zona epiléptica, portanto, otimizando a área ressecada da cirurgia. Para o segundo experimento, os modelos de classificação treinados revelaram precisão satisfatória quando testados em novos dados e em um cenário distinto, demonstrando que é possível utilizar os dados de eletroencefalograma para prever o tipo de displasia cortical focal em pacientes com epilepsia.

**Palavras-chave:** Mineração de dados. Epilepsia. Análise preditiva. Eletroencefalografia. Descargas epileptiformes interictais.

## LIST OF FIGURES

Figure 1 – Methodological Guidelines for the steps 4, 5 and 6.....	22
Figure 2 – The multiple factors that contribute to epilepsy.....	24
Figure 3 – Concept of the Epileptogenic Zone (EZ). ....	27
Figure 4 – The three possible surgery scenarios.....	29
Figure 5 – Treatment steps of a refractory epilepsy patient. ....	31
Figure 6 – Brain wave patterns.....	32
Figure 7 – IED Characteristics. ....	36
Figure 8 – Interictal epileptic discharge (IED) patterns recorded with intracranial electrodes. .....	37
Figure 9 – Steps of KDD. ....	38
Figure 10 – PCA projections and PC calculations. ....	42
Figure 11 – Graphic representation of the $p$ -Value. ....	45
Figure 12 – Illustration of Cohen's $U_3$ ....	47
Figure 13 – Illustration of Hedges' $g$ . ....	49
Figure 14 – Example of $k$ NN classification.....	51
Figure 15 – Example of distributions with different skewness. ....	53
Figure 16 – Notched box plot characteristics. ....	54
Figure 17 – Example of a confusion matrix for two classes. ....	54
Figure 18 – Representation of a ROC curve.....	56
Figure 19 – Scheme of the data mining algorithm. ....	56
Figure 20 – Spike detection algorithm scheme. ....	58
Figure 21 – IED sorting according to their spatial profile (clustering process). ....	60
Figure 22 – Example of cluster selection. ....	65
Figure 23 – Interpretation and evaluation steps.....	66
Figure 24 – Distribution of the samples compared to the normal distribution.....	70
Figure 25 – 3D plot showing the contribution of each feature in principal components. ....	73
Figure 26 – Scatterplot of FCD I and II for the two most significant features.....	74
Figure 27 – Histogram of the outcome groups for awake data. ....	76
Figure 28 – Box plot of the outcome groups for awake data. ....	76
Figure 29 – Histogram of the outcome groups for awake data with threshold. ....	78
Figure 30 – Box plot of the outcome groups for awake data with threshold. ....	78
Figure 31 – Histogram of the outcome groups for sleep data. ....	80

Figure 32 – Box plot of the outcome groups for sleep data. ....	80
Figure 33 – Histogram of the outcome groups for sleep data with threshold. ....	82
Figure 34 – Box plot of the outcome groups for sleep data with threshold. ....	82
Figure 35 – Histogram of the outcome groups for sleep and awake data. ....	84
Figure 36 – Box plot of the outcome groups for sleep and awake data. ....	84
Figure 37 – Histogram of the outcome groups for sleep and awake data with threshold. ....	86
Figure 38 – Box plot of the outcome groups for sleep and awake data with threshold. ....	86
Figure 39 – Confusion matrix and ROC plot for the Ensemble model. ....	89
Figure 40 – Confusion matrix with true positive and false negative rates. ....	90
Figure 41 – Confusion matrix and ROC plot for the refined Ensemble model. ....	92
Figure 42 – Confusion matrix with true positive and false negative rates. ....	92

## LIST OF TABLES

Table 1 – ILAE classification system for FCDs.....	25
Table 2 – Definition of abnormal brain areas.....	26
Table 3 – Pre-operative diagnostic tools. ....	28
Table 4 – Brain waves (EEG bands). ....	33
Table 5 – Patient Characteristics. ....	62
Table 6 – Number of patients in each analysis.....	64
Table 7 – Description of the relevant output variables.....	66
Table 8 – Description of the relevant medical variables. ....	67
Table 9 – Selected features for the classification models.....	72
Table 10 – Selected features for the predictive analysis. ....	74
Table 11 – Resection index statistics - Awake.....	77
Table 12 – Resection index statistics with cluster threshold - Awake.....	79
Table 13 – Resection index statistics – Sleep.....	81
Table 14 – Resection index statistics with cluster threshold – Sleep.....	83
Table 15 – Resection index statistics - Sleep + Awake.....	85
Table 16 – Resection index statistics with cluster threshold - Sleep + Awake.....	87
Table 17 – Ensemble prediction table for new patients.....	90
Table 18 – Ensemble prediction for the FCD III patients, reclassified as dual pathology.....	91
Table 19 – Ensemble prediction table for new patients - Refined model.....	93

## LIST OF ABBREVIATIONS

AED	Anti-Epileptic Drugs
ANN	Artificial Neural Network
CRISP-DM	Cross-Industry Standard Process for Data Mining
CT	Computed Tomography
CTU	Czech Technical University in Prague
DM	Data Mining
EEG	Electroencephalography
EKG	Electrocardiogram
EZ	Epileptogenic Zone
FCD	Focal Cortical Dysplasia
IED	Interictal Epileptiform Discharges
iEEG	Intracranial Electroencephalography
ILAE	International League Against Epilepsy
ISARG	Intracranial Signal Analysis Research Group
IZ	Irritative Zone
KDD	Knowledge Discovery in Data
KNN	$k$ -Nearest Neighbors
MLE	Maximum Likelihood Estimator
MRI	Magnetic Resonance Imaging
NREM	Non-Rapid Eye Movement
PAHO	Pan American Health Organization
PC	Principal Component
PCA	Principal Component Analysis
PET	Positron emission tomography
RI	Resection Index
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SEMMA	Sample, Explore, Modify, Model, Assess
SOZ	Seizure-Onset Zone
SOZI	Seizure Onset Zone Index
SPECT	Single-Photon Emission Computed Tomography
SVDD	Support Vector Data Description
SVM	Support Vector Machine
USD	United States Dollar
WHO	World Health Organization

## CONTENTS

<b>1</b>	<b>INTRODUCTION</b> .....	<b>15</b>
1.1	PROBLEMATICS.....	18
1.2	OBJECTIVES .....	19
<b>1.2.1</b>	<b>General</b> .....	<b>19</b>
<b>1.2.2</b>	<b>Specifics</b> .....	<b>19</b>
1.3	JUSTIFICATION .....	20
1.4	METHODOLOGY .....	21
1.5	WORK STRUCTURE .....	22
<b>2</b>	<b>THEORETICAL FOUNDATION</b> .....	<b>23</b>
2.1	EPILEPSY .....	23
<b>2.1.1</b>	<b>Refractory Epilepsy</b> .....	<b>24</b>
<b>2.1.2</b>	<b>Focal Cortical Dysplasia – FCD</b> .....	<b>25</b>
<b>2.1.3</b>	<b>Epileptogenic Zone – EZ</b> .....	<b>26</b>
<b>2.1.4</b>	<b>Surgery</b> .....	<b>27</b>
<b>2.1.5</b>	<b>Outcomes</b> .....	<b>29</b>
2.2	ELECTROENCEPHALOGRAPHY – EEG .....	32
<b>2.2.1</b>	<b>Sleep and Awake Vigilance Epochs</b> .....	<b>33</b>
<b>2.2.2</b>	<b>Artifacts</b> .....	<b>34</b>
2.3	EPILEPTIFORM DISCHARGES.....	35
<b>2.3.1</b>	<b>Interictal Epileptiform Discharges - IED</b> .....	<b>36</b>
2.4	DATA MINING .....	38
<b>2.4.1</b>	<b>Principal Component Analysis - PCA</b> .....	<b>41</b>
<b>2.4.2</b>	<b><i>k</i>-Means</b> .....	<b>42</b>
<b>2.4.3</b>	<b>Maximal Likelihood Estimation</b> .....	<b>43</b>
<b>2.4.4</b>	<b><i>p</i>-Value</b> .....	<b>43</b>
<b>2.4.5</b>	<b>Wilcoxon rank-sum test</b> .....	<b>45</b>

2.4.6	Effect size statistics.....	46
2.4.7	Classification Models .....	50
2.4.8	Descriptive Statistics .....	52
2.4.9	Statistical Visualization.....	53
2.5	EEG ANALYSIS TECHNIQUE.....	56
3	<b>METHODS.....</b>	<b>61</b>
3.1	PRE-PROCESSING .....	61
3.2	DATA AND PATIENTS .....	62
3.3	DATA MINING .....	64
3.4	INTERPRETATION AND EVALUATION .....	66
3.4.1	Treatment of Zeros.....	67
3.4.2	Calculation of Outliers.....	67
3.4.3	Calculation of the Resection Index - RI.....	68
3.4.4	Statistical Hypothesis .....	69
3.4.5	Definition of a Threshold.....	70
3.4.6	Statistics and Graphs .....	70
3.5	PREDICTIVE ANALYSIS .....	71
4	<b>RESULTS AND DISCUSSION.....</b>	<b>75</b>
4.1	AWAKE .....	75
4.2	SLEEP .....	79
4.3	SLEEP AND AWAKE.....	83
4.4	DISCUSSION OF THE STATISTICAL RESULTS .....	88
4.5	CLASSIFICATION RESULTS .....	88
4.5.1	Ensemble (Subspace $k$ NN).....	89
4.5.2	Ensemble (Subspace $k$ NN) - Refined .....	91
4.6	DISCUSSION OF CLASSIFICATION RESULTS .....	93
5	<b>CONCLUSION.....</b>	<b>94</b>
	<b>REFERENCES .....</b>	<b>95</b>



## 1 INTRODUCTION

Epilepsy is a group of chronic neurological diseases that affects between 0.5% to 1% of the world's population (BANERJEE; HAUSER, 2008). It is characterized by recurrent unprovoked seizures, which can be described as a temporary change in brain functions, triggered by abnormal electrical discharges within the human brain (FISHER et al., 2005). The seizures can vary from brief lapses of awareness, involuntary muscle movements, which can be partial or generalized, sensation disturbances, to severe and prolonged convulsions. Seizures also vary in frequency, from once a year to several per day (WHO, 2019). There are a couple of definitions of epilepsy, formal or practical (FISHER et al., 2005), the more widespread one is that epilepsy is a condition characterized by two or more unprovoked seizures (BANERJEE; HAUSER, 2008).

The disease can cause several consequences, neurobiological, cognitive, psychological and social to someone's life, affecting its quality of life negatively (FISHER et al., 2014). It also has social and economic impacts since it can be associated with higher mortality rates, cognitive deficits, loss of productivity, and accidents (KERR, 2012). Despite affecting people of all ages, approximately 75% of the cases begin during childhood (STAFSTROM; CARMANT, 2015).

According to the World Health Organization (WHO), more than 50 million people worldwide have epilepsy, making it one of the most common neurological diseases globally (2019). About 80% of them live in low- and middle-income countries. It is estimated that there are currently around 8 million people with epilepsy in Latin America alone, and approximately 2 million in Brazil. About 50% of them do not receive appropriate diagnoses, three-quarters do not receive proper treatments and less than 50% have access to anti-seizure medications. Additionally, two-thirds of the countries in the region do not have a specific program for the care of people with epilepsy, and 80% of them do not have proper legislation about the disease (PAHO, 2013).

Brazil follows its Latin American neighbors, with a higher epilepsy rate when compared to the developed countries, the phenomenon was researched by several studies, both in urban and rural areas (SIQUEIRA et al., 2016). This can be attributed to worse social-economic conditions in these countries, which leads to inadequate or lack of treatment in the

case of neuroinfections (BRUNO et al., 2013), also deficient prenatal care and subpar childbirth conditions (NUNES et al., 2011).

Treatment options include dietary therapies and pharmacotherapy, but approximately one-third of the patients can be or become pharmacoresistant, also known as refractory. The sub-group of refractory epilepsy represents patients with lesional epilepsy that can profit from an epileptosurgery, which is a surgery to remove the epileptogenic brain tissue (TANG et al., 2017).

Epilepsy can be either congenital or acquired, with the former being caused due to malformation of the brain before birth, and the latter being caused by several factors, including head injuries, infections, and brain tumors (WHO, 2019) (SYNAPSE, 2019). It can also be classified by its onset as generalized, focal, or unknown. The generalized type is mostly genetically determined, and it affects both brain hemispheres. People with this kind of seizure display impaired awareness during an episode, in addition to the motor and non-motor symptoms. Differently, focal types begin in a localized part of the brain, and its clinical manifestations depend on the area of the brain and the propagation of the epileptogenic discharges. Usually, people are aware or partially aware during this kind of seizure. Finally, the unknown type is used to classify an undefined onset, and it is usually reclassified as generalized or focal when further information from exams is provided (FISHER et al., 2017).

Malformations of cortical development are one of the causes of medically refractory epilepsy, the Focal Cortical Dysplasia (FCD) being one of them. These are structural abnormalities in the cerebral cortex development during early intrauterine life. Both genetical and acquired factors can be involved in the development of FCD, and there are several proposed classifications to these structural abnormalities, but in general, three types of cortical dysplasia are recognized (BAE et al., 2012). These type differences are characterized by morphological and symptomatic variations and can be detected by diagnostic imaging techniques such as Magnetic Resonance Imaging (MRI) and Electroencephalography (EEG). (KABAT; KRÓL, 2012).

EEG is an electrophysiological test used to evaluate and record electrical activity in the brain. This technique usually measures electrical activity using a set of electrodes on the head surface (scalp). The standard scalp EEG allows the measurement of brain activity from lateral parts of the brain with limited spatial resolution. If the source area of the seizures is

unclear by the imaging techniques, the intracranial EEG (iEEG) is required. iEEG measures local field potentials, and it is usually used to help the diagnostics on medically intractable epilepsy patients as a tool to find and define the epileptogenic zone (KOVAC et al., 2017) (LACHAUX et al., 2003). For epilepsy, the epileptiform EEG recordings can be divided into two primary time epochs: during seizures (ictal) and the activity between seizures (interictal) (DEWOLFE; MALOW, 2012).

EEG and iEEG recordings poses a significant challenge to data analysis because of the large amount of data generated, as well as its intrinsic characteristics such as the noise, due to background brain activity and from other biological signals, spatial and temporal components, and the fact that a proper analysis requires a wide range of data mining and statistical techniques (FLEXER, 2000).

Data mining (DM) is the study of gaining useful insights and patterns from data (AGGARWAL, 2015). It is an interdisciplinary area closely related to statistics, information science, machine learning, among others. Data mining is usually considered a fundamental step of the Knowledge Discovery in Databases (KDD). The entire KDD process can be described by several steps applied to the data: cleaning, selection, transformation, mining, pattern evaluation, and knowledge presentation (HAN et al., 2011).

In the last decades, companies, public institutions, and laboratories are generating an ever-greater amount of data. This data is usually stored in an attempt to discover useful implicit information that will help them with planning, decision making, market analysis, or boosting their productivity.

With the advent of the automated systems and sensors, the amount of data generated and needed to be stored easily reaches the order of gigabytes, terabytes or petabytes (BERRY; LINOFF, 2004). This data is, therefore, available to analysis, and with the computational power becoming more affordable and commercial advanced DM software and tools becoming broadly available, the field is growing both financially and in importance (LEVENTHAL, 2010).

According to a new market research report, the market for DM tools is expected to grow from USD 591.2 Million in 2018 to USD 1,039.1 Million by 2023, an 11.9% growth over year during the period (MARKETS AND MARKETS, 2018), which represents an increasing demand and awareness about the importance of the field.

The area of bioinformatics can also benefit greatly from DM. Especially after the recent progress in biology, medical science, and biotechnology, the bioinformatics has become a data-intensive field, producing large amounts of data that require in-depth analysis. The efficient and scalable methods for mining interesting patterns, and visualization techniques, allied to the complexity of biological data, make the combination of the two fields exciting and challenging.

## 1.1 PROBLEMATICS

The problem of defining the epileptogenic structure in the brain is a complex and multidisciplinary challenge involving the understanding of the complexity of the brain structures and functioning.

From the medical-biological area, we have the formal definition of the epileptogenic zone (EZ) as “the minimum amount of cortex that must be resected (inactivated or completely disconnected) to produce seizure freedom” as well as other important structures that characterizes it like the seizure-onset zone (SOZ), irritative zone, symptomatogenic zone, epileptogenic lesion and the functional deficit zone (LÜDERS et al., 2006) (ROSENOW; LÜDERS, 2001). Even though there are some diagnostic tools, such as MRI and EEG, for helping the definition of these cortical zones, for the correct acquisition, pre-processing, and analysis of the data from these techniques, the bioinformatics field must be involved.

Signal analysis techniques for the correct detection of the EZ are being researched and developed for at least the last ten years, after all, this is a crucial objective for achieving a high success rate in an epileptosurgery. The current understanding of the area, in opposition to the old belief, is that the activity is not focalized, but instead, is the interaction of complex brain networks (BARTOLOMEI et al., 2017).

In order to understand and model these networks, diverse types of approaches are being proposed, from graph analysis (WILKE et al., 2011) to machine learning algorithms (GRINENKO et al., 2018), with varied results. Recently, an approach of mapping the network organization of the brain in a modular way, representing it by multiple sub-regions (clusters) with different propagation trajectories, showed that for 12 of 14 cases the most active sub-region is localized within the seizure onset zone (JANCA et al., 2018). Improving this algorithm

would prove to be difficult, especially considering its complexity and since it is a result of years of development. Although it seems that the clustering process could be tuned, the most critical aspect lacking is the validation and test of the algorithm for a higher number of patients followed by statistical analysis.

Knowing that the algorithm proposed by Janca et al. (2018) has promising results, and opens space for derived works, the question of this work is defined as the following: **“How to extract representative parameters from the results of the chosen method, when applied to FCD epilepsy patient data, to help the better delineation of the EZ?”**.

## 1.2 OBJECTIVES

This section includes descriptions of the general and the specific objectives of this work.

### 1.2.1 General

Identify and verify representative parameters of epileptic activity in multichannel iEEG recordings, using data mining techniques, in order to define the critical brain epileptogenic structure.

### 1.2.2 Specifics

- Verify the statistical significance of the chosen algorithm for the selected data, comparing its results with the patient’s surgery outcome results, with the help of an index.
- Verify the differences between sleep and awake vigilance epochs testing its statistical significance for outcome prediction, using an index.
- Find an optimal threshold with the most significant statistical difference between groups.
- Discover and test the effectiveness of the chosen algorithm’s outputs for detecting the FCD type (I, II), using predictive analysis methods.

### 1.3 JUSTIFICATION

Researching a topic related to the medical area is always a difficult decision, there are people involved, and there is a concern about personal data, especially when it is disease-related. However, it is also quite rewarding, since the smallest effort, alongside efforts from others, can benefit a lot of people and make a difference in someone's life.

Epilepsy is the kind of disease that most people think they know about, but unfortunately, there are a lot of misconceptions coming from the general public. It is also a disease that could be easily avoided in a lot of cases, but due to the "treatment gap", particularly in undeveloped countries, people are developing at unusual higher rates.

On the other hand, refractory epilepsy, also known as uncontrolled or recurring, has a difficult treatment. Despite advances both in medicine and imaging tools, the rates of seizure-free outcomes after surgery have not improved as much, and even after eliminating the intractable cases, questions like "How much tissue must be resected to obtain a seizure-free outcome without compromise of memory and other brain functions?" are still to be answered (FEINDEL et al., 2009).

It is also essential to bring that, compared to other neurological conditions, epilepsy is lagging in terms of funding and research, and diseases like Alzheimer's and Parkinson have higher investment rates given the incidence rates (GRABOWSKI et al., 2018).

From a technological perspective, the field of data mining is becoming ever more critical, mainly due to the flexibility of its uses. Areas with a massive demand for data mining include science, business, industry, and web, among others. Almost any field with a good amount of data can benefit from it, and with the increasingly higher usage of computers and automatization, the demand is also growing rapidly.

As Frawley (1992) tragically presents, "Computers have promised us a fountain of wisdom but delivered a flood of data.". Moreover, to get to this fountain of wisdom, we need to go through a non-trivial process that involves a good amount of techniques and diverse types of data. This is what makes DM so challenging, but also really satisfying to work with.

## 1.4 METHODOLOGY

This section is written with the purpose of describing and explaining the methodology deployed in this study in order to achieve its objectives.

1. Research: Do an exploratory bibliographical research focusing on data science and refractory epilepsy.
2. Data Collection and Selection: iEEG patient data were recorded in Motol University Hospital of Charles University in co-operation with CTU in Prague, Czech Republic. Data collection was approved by the institutional ethics committee, and informed personal or parental consent was obtained. Dataset consists of long term iEEG recordings, clinical evaluation of zones and after-surgery outcome.
3. Processing: The data was processed using the chosen algorithm. The entire process is composed of several steps:
  - a. Detection of the interictal epileptical discharges (IED) on the raw iEEG data using statistical signal models (JANCA et al., 2015). Followed by the calculation of statistics.
  - b. Clusterization of the detected IED, according to patterns of onset and electrical field of propagation (JANCA et al., 2018). A high separability coefficient setting was chosen for all patients due to the better results provided in previous tests (INÁCIO; JANCA, 2019).
  - c. Manual selection of the clusters discarding the ones with evident artifacts or false positives.
  - d. Calculation of the cluster statistics.
4. Statistics: Calculation of indexes and statistical significance in comparison with surgery outcome information, the same for “awake” and “sleep” vigilance epochs separately and calculation of an optimal threshold.
5. Exploration: Further analysis using data mining techniques to discover and prove the output data effectiveness in predicting the patient’s FCD types.
6. Results: Evaluation and discussion of the results.

Figure 1 details further methodological aspects for the post-processing steps.

Figure 1 – Methodological Guidelines for the steps 4, 5 and 6.



Source: Adapted from Dresch (2015)

## 1.5 WORK STRUCTURE

This work is structured into five chapters.

- **Chapter 1** introduces the subject and all the involved topics, explains the problematics and describes the work's methodology.
- **Chapter 2** introduces the theoretical foundations necessary for the understanding of this study, both for the medical and the technical aspects.
- **Chapter 3** describes the methods and execution of the analysis. First for the statistical hypothesis evaluation and second for the predictive analysis.
- **Chapter 4** presents and discusses the obtained results. It also details similarities and differences from related studies.
- **Chapter 5** highlights the findings of this work and discusses future opportunities.



## 2 THEORETICAL FOUNDATION

### 2.1 EPILEPSY

This chapter presents a brief introduction about epilepsy and brings up important concepts about the disease that are necessary to the understanding of this work.

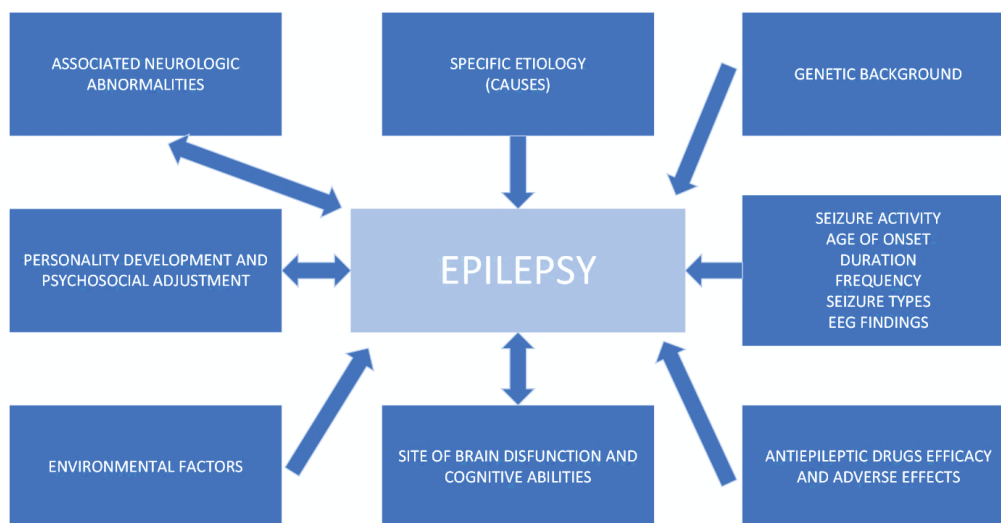
Epilepsy is a group of chronic diseases characterized by recurrent unprovoked uncontrolled electrical activity in the brain. An occurrence of abnormally excessive or synchronous electrical discharges within the human brain is called “seizure” (FISHER et al., 2005).

Seizures trigger temporary changes in the brain functions and may have several effects, varying from brief lapses of awareness, involuntary muscle movements, sensation disturbances to severe and prolonged convulsions. They are often related to brain injuries or genetic factors, such as hereditary predisposition, but also due to unknown factors (WHO, 2019).

Seizures vary in frequency, from once a year to several per day, and people are considered having epilepsy if they have two or more unprovoked seizures more than 24 hours apart.

Epilepsy is not a single disorder but rather a broad group of conditions altering the brain functions, causing a variety of pathologic processes (ENGEL; PEDLEY, 2008). As a heterogeneous condition multiple and complex interacting factors can contribute to the totality of the disease in one patient, as seen in Figure 2.

Figure 2 – The multiple factors that contribute to epilepsy.



Source: Extracted from Engel and Pedley (2008).

### 2.1.1 Refractory Epilepsy

Uncontrolled seizures are a considerable burden and have a negative impact on the health, cognitive, psychological, and social aspects of someone's life. The negative aspects include physical injuries, accidents, depression and anxiety, deficits in memory and thinking skills, developmental delays in children, and sudden death (SIRVEN; SHAFER, 2014).

Some patients have a condition where the seizures cannot be controlled with anticonvulsant medicaments, also known as anti-epileptic drugs (AED). This condition is often called “uncontrolled”, “intractable” or “refractory” epilepsy.

For these patients, which represent about one-third of the epilepsy cases, seizure freedom is very unlikely to be achieved with further manipulation of AED therapy (SIRVEN; SHAFER, 2014).

Although the concept of drug resistance may appear intuitive, a precise definition is vital to improve patient care and facilitate clinical research and diagnostics. Like many topics in science, achieving a consensus is a hard task, so authors and researchers usually used different recommendations.

In response to this, in 2009, a multidisciplinary task force of the International League Against Epilepsy (ILAE) decided to propose a consensus definition of drug-resistant epilepsy. The following description was proposed:

Drug-resistant epilepsy may be defined as failure of adequate trials of two tolerated and appropriately chosen and used AED schedules (whether as monotherapies or in combination) to achieve sustained seizure freedom. (KWAN et al., 2010)

Patients should only be considered refractory if wrong diagnosis or suboptimal treatment trials are out of the question. If the appropriate response is not achieved with the AED trials patients often need further diagnostic testing that includes video EEG monitoring or additional studies and diagnostic techniques. Patients with refractory epilepsy are the ones most likely to be recommended to undergo brain surgery (JEROME ENGEL, 2014).

### 2.1.2 Focal Cortical Dysplasia – FCD

Focal cortical dysplasia (FCD) is a malformation of cortical development. It is responsible for a large share of the refractory epilepsy cases being the most common cause in the pediatric population and the second or third most common in adults (KABAT; KRÓL, 2012).

These malformations are structural abnormalities in the cerebral cortex development during early intrauterine life and can be caused by genetic or acquired factors.

There are several proposed classifications to these structural abnormalities, but in general, three types of cortical dysplasia are recognized (I, II, and III) distinguishing the types by its form and association with another kind of lesion (BAE et al., 2012). The proposed ILAE classification system for focal cortical dysplasia, with the description of the cortex abnormalities, can be seen in Table 1.

Table 1 – ILAE classification system for FCDs.

<b>FCD Type I (isolated)</b>	With abnormal radial cortical lamination (FCD Ia)	With abnormal tangential cortical lamination (FCD Ib)	With abnormal radial and tangential cortical lamination (FCD Ic)	
<b>FCD Type II (isolated)</b>	With dysmorphic neurons (FCD IIa)		With dysmorphic neurons and balloon cells (FCD IIb)	
<b>FCD Type III (associated with principal lesion)</b>	Associated with hippocampal sclerosis (FCD IIIa)	Adjacent to a glial or glioneuronal tumor (FCD IIIb)	Adjacent to vascular malformation (FCD IIIc)	Adjacent to any other lesion acquired during early life (FCD IIId)

Source: Adapted from Blümcke et al. (2011).

For FCD III, distinguishing between dual pathology and type III may be unclear, so commonly they might also be classified as FCD I or II with associated pathology (BAE et al., 2012).

The seizures in FCD are not easy to control with pharmacological treatment, and thus often, patients are classified as refractory. Therefore, the surgical procedure has been a crucial alternative treatment for these patients (KABAT; KRÓL, 2012).

### 2.1.3 Epileptogenic Zone – EZ

The formal definition of the epileptogenic zone (EZ) according to Lüders and Najm (2006) reads as “the minimum amount of cortex that must be resected (inactivated or completely disconnected) to produce seizure freedom”. In their work, they also describe other important structures that characterize it like the seizure-onset zone (SOZ), irritative zone, symptomatogenic zone, epileptogenic lesion, and the functional deficit zone. The description of the abnormal brain structures and the tools used to diagnose them can be seen in Table 2.

Table 2 – Definition of abnormal brain areas.

<b>Brain Area</b>	<b>Definition</b>	<b>Measure</b>
Irritative Zone (IZ)	Area of cortex that generates interictal spikes.	EEG
Seizure Onset Zone (SOZ)	Area of cortex that initiates or generates seizures.	EEG
Epileptogenic lesion	Structural pathology of the brain that is the direct cause of seizures.	CT, MRI, tissue pathology
Symptomatogenic zone	The portion of the brain that produces the first clinical symptoms.	EEG, behavioral observation
Functional deficit zone	Cortical area producing nonepileptic dysfunction.	Neurologic exam, neuropsychology PET, SPECT
Epileptogenic zone (EZ)	The total area of the brain that is necessary to generate seizures and must be removed to abolish seizures.	Unknown
<i>CT – Computed Tomography; EEG – Electroencephalography; PET – Positron emission tomography; SPECT – Single-photon emission computed tomography.</i>		

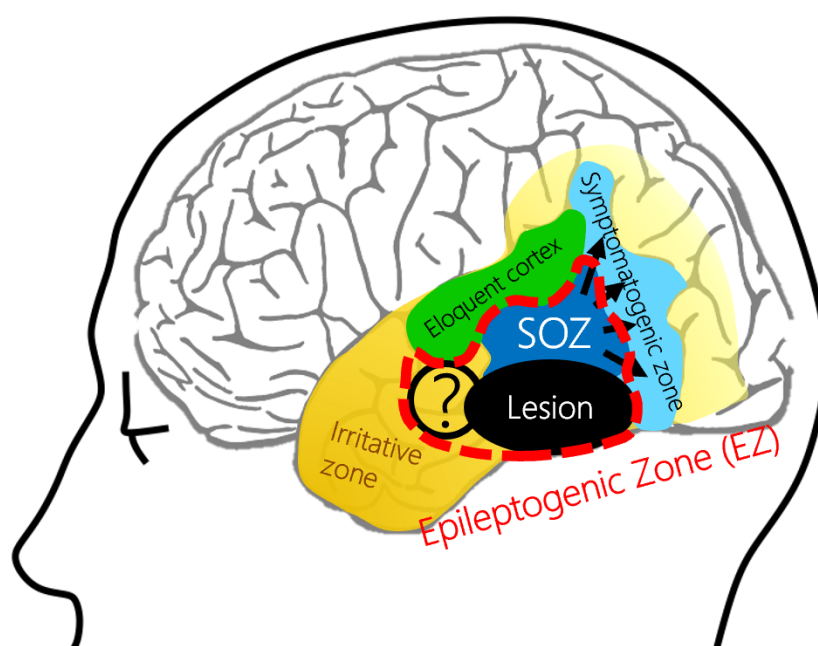
Source: From Engel and Pedley (2008).

The complex brain epileptic network is simplified to the concept of epileptic zones that mutually overlap, which is represented in Figure 3. A part of the lesion that initiates seizures

is called the seizure onset zone, together with a highly irritative tissue, represents the epileptogenic zone, whose complete removal/disconnection leads to seizure-freedom.

Surrounding connected structures irritates from the EZ and produce interictal epileptiform discharges, that are widely spreading within the irritative zone and often cover the whole epileptic network. More detailed stratification of the epileptic network defines functionally specific symptomatogenic zones that are responsible for seizure semiology. (LÜDERS et al., 2006).

Figure 3 – Concept of the Epileptogenic Zone (EZ).



Source: Translated from JANCA, R. class presentation (2019)

### 2.1.4 Surgery

Epilepsy surgery is a medical procedure that removes or “disconnects” an area of the brain where the seizures originate. It has better results when the seizures always originate in a single part of the brain.

Surgery is hardly the first option of treatment due to its inherent risks, and it is only done after a series of procedures and pre-surgical assessments to define whether the patient is eligible for the surgery (RYVLIN; RHEIMS, 2008).

There are several types of epilepsy surgeries. The type depends mainly on the location of the neurons that trigger the seizure and the age of the patient. For the sake of concision and objectiveness, only the one relevant to the analyzed patients is going to be described.

Resective surgery is the most common epilepsy surgery, and it is done by a resection (removal) of a small portion of the brain. Brain tissues are removed from the area where seizures originate (SOZ) or an area that helps to spread the seizure activity (IZ), usually, the site of a tumor, lesion, or malformation. The conjunct of these areas is considered as the epileptogenic zone (EZ).

Before undergoing brain surgery, extensive testing is done to locate the abnormal areas in the brain defined by the EZ, and to ensure that removing the region of the brain will not impact speech, mobility or quality of life, which are functions of the eloquent cortex. Additionally, other treatment options are evaluated, previous treatments are reviewed and social and health aspects that impact the patient's life are determined (KIRIAKOPOULOS, 2018).

The standard procedures to evaluate the patients and to detect and map the source of the abnormal brain activity are described in Table 3. The referred additional tests are recommended when the origin of the seizures is still unclear.

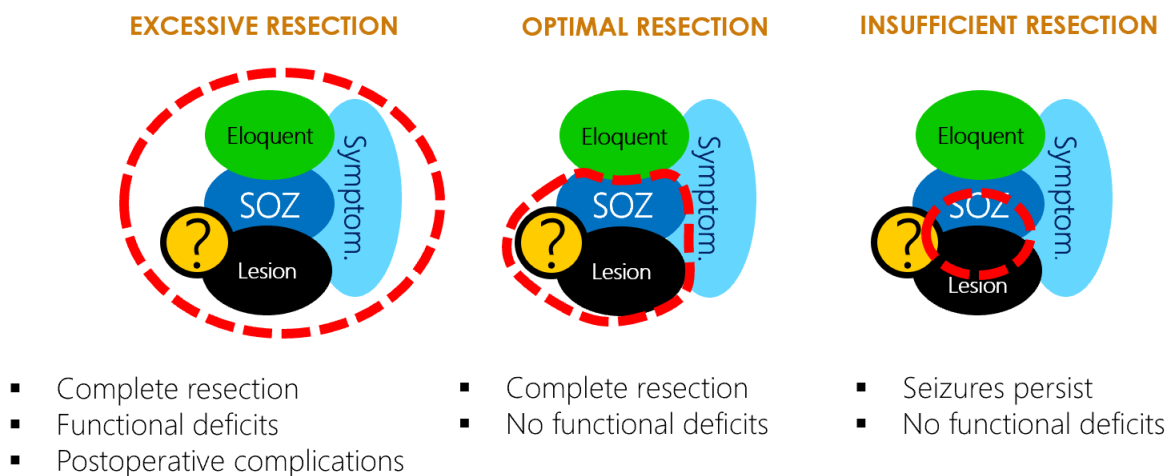
Table 3 – Pre-operative diagnostic tools.

<b>Basic diagnostic tools</b>	<b>Description</b>
Scalp electroencephalogram (EEG)	Electrodes placed on the scalp to measure brain activity. Detected patterns can suggest the affected brain area.
Video EEG	Continuous EEG with a video-monitoring test in a hospital. Correlation between EEG activity and the symptoms during the seizure help to define the area of the brain where the seizure starts.
Magnetic resonance imaging (MRI)	Imaging test used to detect brain abnormalities that are causing seizures, such as tumors or damaged areas.
<b>Additional Tests</b>	<b>Description</b>
Invasive EEG monitoring (iEEG)	If scalp EEG fails to detect the seizure-inducing area, surgically placed electrodes are put on the brain surface or implanted in the brain. Measures are done while the patient is unconscious.
Video EEG with invasive electrodes	After the implantation surgery, the video and EEG data are captured during a defined time, while the patient is in hospital and not taking medications.

Source: Adapted from Mayo Clinic (2019)

There are three possible scenarios regarding the surgery resected area and its effects on patient's health (Figure 4). An excessive resection happens when the resected area is larger than the EZ, reaching areas of the eloquent cortex; An optimal resection happens when none but the areas that comprehend the epileptogenic zone (EZ) are resected; And an insufficient resection occurs when the resected area is smaller than the necessary. Each one of these scenarios produce impacts on the health and the surgery outcomes of the patient.

Figure 4 – The three possible surgery scenarios.



Source: Translated from JANCA, R. class presentation (2019)

### 2.1.5 Outcomes

As a system of classification of postoperative outcomes, Jerome Engel proposed the following scheme that has become a standard when reporting the results in the epilepsy medical literature, it reads as follows (ENGEL, 1993):

**Class I:** Seizure free or no more than a few early, nondisabling seizures; or seizures upon drug withdrawal only.

**Class II:** Disabling seizures rarely occur during a period of at least two years; nocturnal seizures.

**Class III:** Worthwhile improvement; seizure reduction for prolonged periods but less than two years.

**Class IV:** No worthwhile improvement; some decrease, no decrease, or worsening are possible.

In 2001 the International League Against Epilepsy (ILAE) also proposed a new classification scheme. The goal was to introduce a more objective classification (WIESER et al., 2001). In this work, though, the Engel classification system was chosen due to its standard status.

Factors that contribute to a favorable seizure-free outcome include:

- Absent or infrequent secondarily generalized convulsions;
- A lesion in a well-delimited area;
- No or minimum overlap with regions of the eloquent cortex;
- The absence of diffuse pathology;
- Complete epileptogenic zone resection;
- Type of pathology.

Additionally, a seizure-free outcome after two years foresees a long-term outcome at five years or more (PASSARO; BENBADIS, 2018).

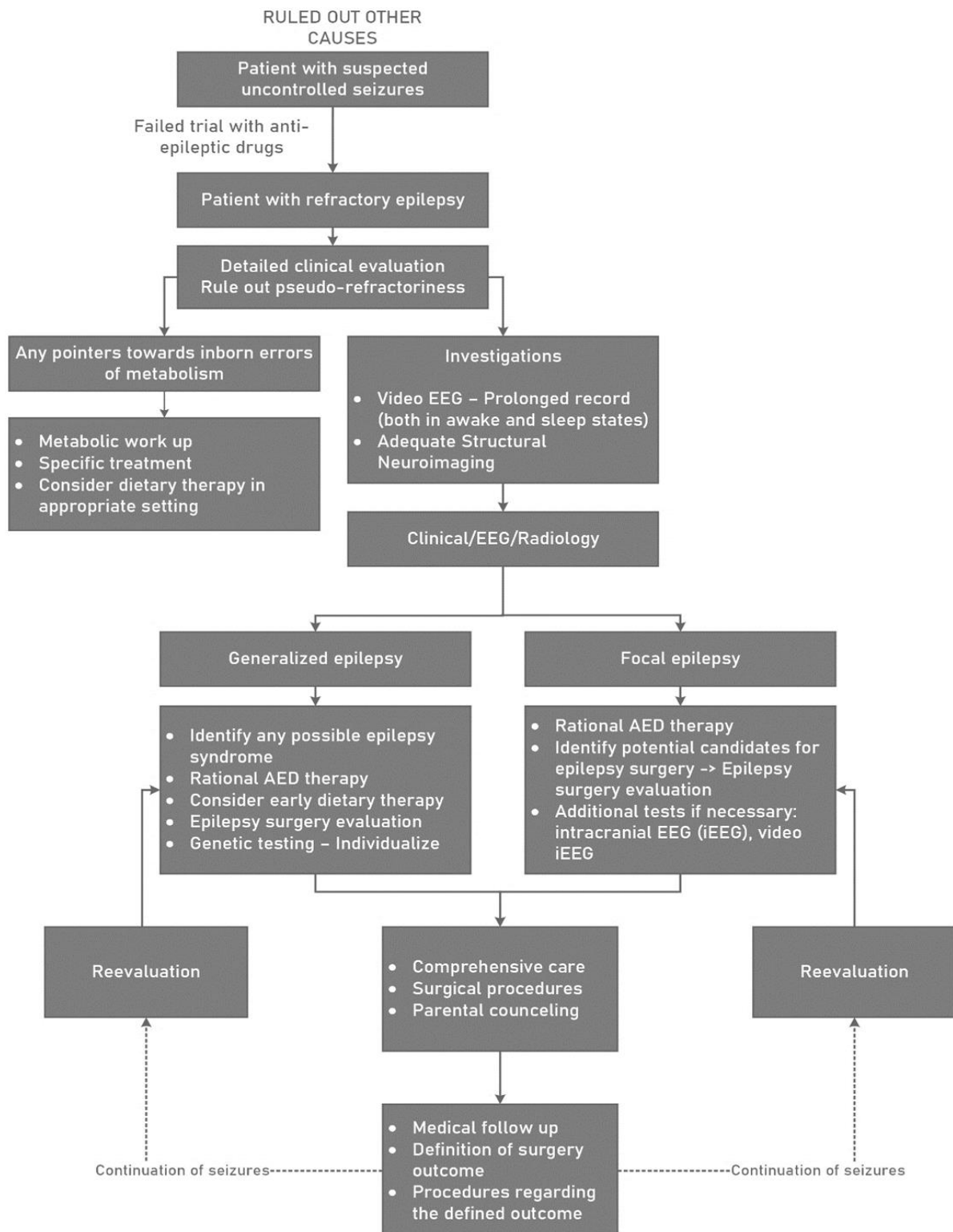
The seizure-free outcome rates range from 20-90%, depending on the type of lesion, with some having higher favorable outcome rates than others (GUAN et al., 2016).

For focal cortical dysplasia (FCD), several studies show a short-term follow-up with seizure-free rates of 40–86%. Long term follow-up studies seem to be a bit rarer, but these show a favorable outcome usually greater than 60% (KRAL et al., 2007) (FAUSER et al., 2015).

The steps of a refractory epilepsy patient from the diagnosis to the definition of the surgery outcome are represented by the diagram in Figure 5.



Figure 5 – Treatment steps of a refractory epilepsy patient.



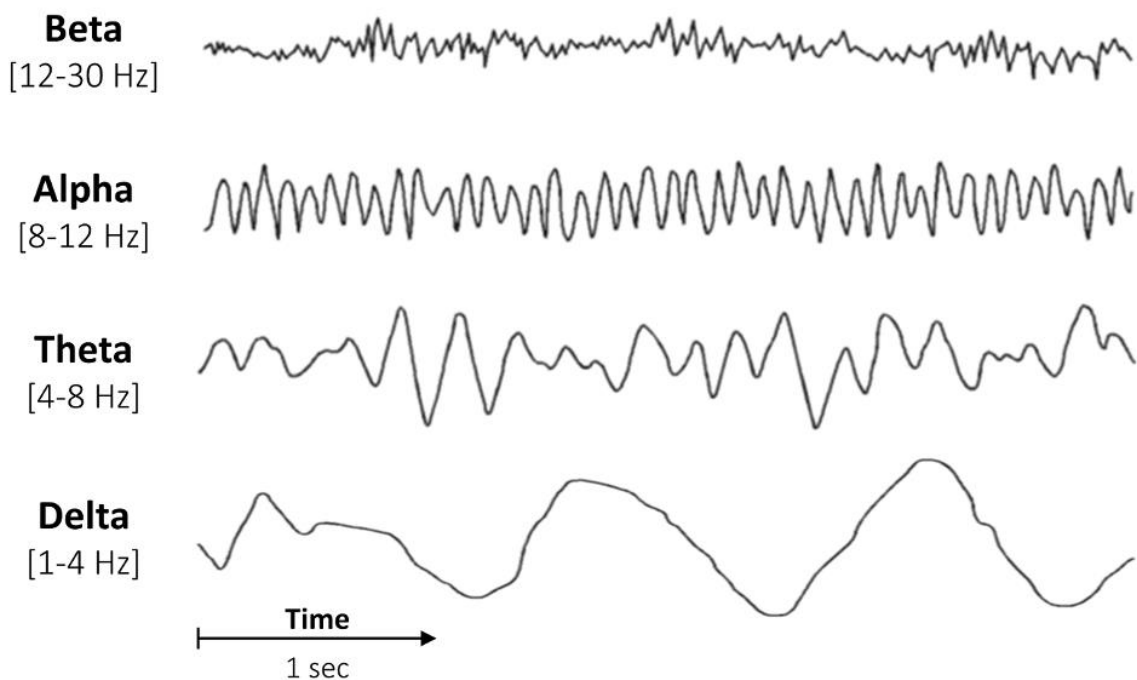
Source: Adapted from Aneja and Jain (2014)

## 2.2 ELECTROENCEPHALOGRAPHY – EEG

Electroencephalogram (EEG) is an electrophysiological test that records the electrical activity of the brain. It tracks and records the brain waves using metal electrodes placed on the scalp (EEG), on the exposed surface of the brain, or using depth probes inserted in the brain (iEEG). The electrodes analyze the electrical activity of the area and send the results to a computer that records the results.

The electrical impulses recorded from the electrodes reflect the cortical activity and appear in the image activity as wavy lines. These patterns allow doctors to detect abnormal activity such as seizures and other brain disorders. It is one of the most important tests to help with the correct diagnosis and treatment of epilepsy. Figure 6 shows the examples of the brain wave patterns that can be detected with the EEG.

Figure 6 – Brain wave patterns.



Source: From Vallat (2018)

The electrodes in the standard EEG are placed according to a standard known as 10/20 system, and in the intracranial EEG (iEEG), the placement of the electrodes depends on the affected brain region and is defined after exams using diagnostic tools to calculate the precise

locations. The scheme of arrangement of the electrodes is referred to as “montage” and can be either a bipolar montage or a referential one. The bipolar uses two electrodes for one channel, having a reference electrode for each channel. As for the referential montage, just one common reference electrode serves to all channels.

It is essential to differentiate the electrodes from the channels. The electrodes are a single point of contact between the acquisition system and the brain area. They are labeled with a letter and a number that refer to its placement on the brain, and they can also be organized as arrays, grids, strips, and probes. The channels, however, are the regular measurements of the potential difference between two electrodes, after an analog-to-digital conversion, which results in a signal represented as a time series in the data. The channels can be changed to represent different types of montages (THE MCGILL PHYSIOLOGY VIRTUAL LABORATORY, 2005).

The detected brain waves are analyzed into two components, amplitude, and frequency. The amplitude represents its electrical strength; they are quite small and are measured in  $\mu\text{V}$ . As for the frequencies, it depends on the detected brain waves.

The brain waves can be categorized into four primary groups: alpha, beta, theta, and delta, and each one has specific frequencies and characteristics (Table 4).

Table 4 – Brain waves (EEG bands).

Wave Group	Frequency	Characteristics
<b>Beta</b>	> 12 Hz	It is generally regarded as a normal rhythm. They are closely linked to motor behavior and are predominant during states of alertness, anxiety, or with eyes open.
<b>Alpha</b>	8 – 12 Hz	They are predominant during wakeful relaxation with closed eyes and are reduced with open eyes, drowsiness, and sleep.
<b>Theta</b>	4 – 8 Hz	Cortical theta is observed frequently in young children. In adults, it tends to appear during meditative, drowsy, hypnotic, or sleeping states, but not during the deepest stages of sleep.
<b>Delta</b>	0.5 – 4 Hz	They are usually associated with deep sleep stages.

Source: Adapted from The McGill Physiology Virtual Laboratory (2005)

### 2.2.1 Sleep and Awake Vigilance Epochs

EEG is sensitive to a continuum of states ranging from stress, alertness, resting state, hypnosis, and sleep (TEPLAN, 2002). These are caused by changes in the dominant brain

waves in action. Epileptic seizures are also strongly influenced by the sleep-wake cycle, as evidenced by the occurrence of the seizures. In as many as one-third of patients with partial epilepsy, seizures may not be present during wakefulness, but only during sleep, despite that, most have both daytime and night-time seizures.

During sleep, there are many shifts of states within the brain, the so-called sleep stages. These changes of state are thought to influence the brain's epileptic activity in people with epilepsy. Some seizures occur predominantly at a particular stage of sleep (EPILEPSY ACTION AUSTRALIA, 2017). These variations of activity in the brain when moving between the different stages of sleep, and between sleep and awakening states may show more unusual electrical activity or activations in different regions and can be essential for the detection of interictal activity helping the diagnosis, particularly, if the waking recording was normal. Awake recordings are often obscured by muscle and movement artifacts, particularly in children and adults who are unable to cooperate or relax during the recordings. Thus, EEGs in patients with suspected seizures should always include sleep, although the actual sleep recording generally does not have to exceed 30 minutes (SHAMSAEI, 2014).

### **2.2.2 Artifacts**

When dealing with EEG data, there are signal distortions that can be observed and are typical among the recorded data that are not generated by the brain; these are called artifacts.

An artifact is usually a sequence with higher amplitude and different shape in comparison to signal sequences. Some artifacts may mimic actual epileptiform abnormalities or seizures and may cause confusion even to experts (SAZGAR; YOUNG, 2019).

The artifacts in the recorded EEG can be classified as patient-related (physiologic) or technical (non-physiologic). The patient-related artifacts are unwanted physiological (biological) signals that can disturb the EEG significantly. Technical artifacts are usually from electrical phenomena or devices in the recording environment.

The most common EEG artifact sources can be classified in the following way (TEPLAN, 2002):

#### Physiologic:

- Any minor body movements (such as tongue movement, or swallowing)
- Muscular contractions
- Cardiac rhythm (EKG)
- Eye movements
- Sweating

#### Non-physiologic:

- 50/60 Hz
- Impedance fluctuation
- Cable movements
- Broken wire contacts
- Electrode popping
- Too much electrode paste/jelly or dried pieces
- Low battery

Filtering out the artifact segments from the EEG traces can be managed by the trained experts or automatically. For better discrimination of different physiological artifacts, additional electrodes for monitoring eye movement, electrocardiogram (EKG), and muscle activity may be necessary.

### 2.3 EPILEPTIFORM DISCHARGES

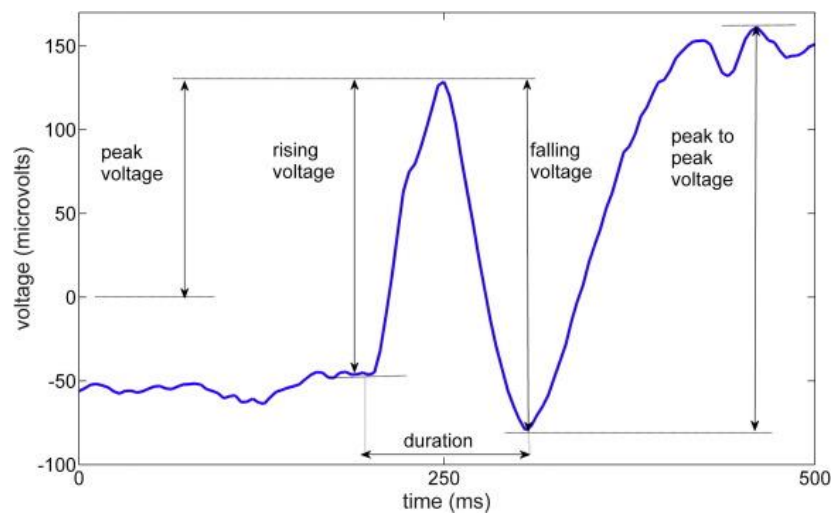
Epileptiform discharges are the intense electrical activity that is generated during (ictal) or in between seizures (interictal). They have attributes and patterns that permit their detection among the EEG data. The seizures are infrequent events in most patients, despite some strategies to stimulate seizures, like medication or sleep deprivation, making the recording of ictal EEG time-consuming, labor-intensive, and luck dependent. The pillar of diagnosis, therefore, lies in the detection of the interictal epileptiform discharges (IEDs), the “in-between” seizures activities. The epileptiform discharges are transients with a characteristic "spiky"

morphology and are commonly referred as spikes. In the diagnosis of epilepsy and localization of the seizure onset zone, both the interictal and ictal recordings are extremely informative.

### 2.3.1 Interictal Epileptiform Discharges - IED

The International Federation of Societies for Electroencephalography and Clinical Neurophysiology describes interictal discharges as a subcategory of "epileptiform pattern" which is defined as "distinctive waves or complexes, distinguished from background activity, and resembling those recorded in a proportion of human subjects suffering from epileptic disorders..." (NOACHTAR et al., 1999). This definition is somewhat circular and unclear, meaning the description is based on experience and also can explain the difficulty of agreement between specialists when identifying IEDs.

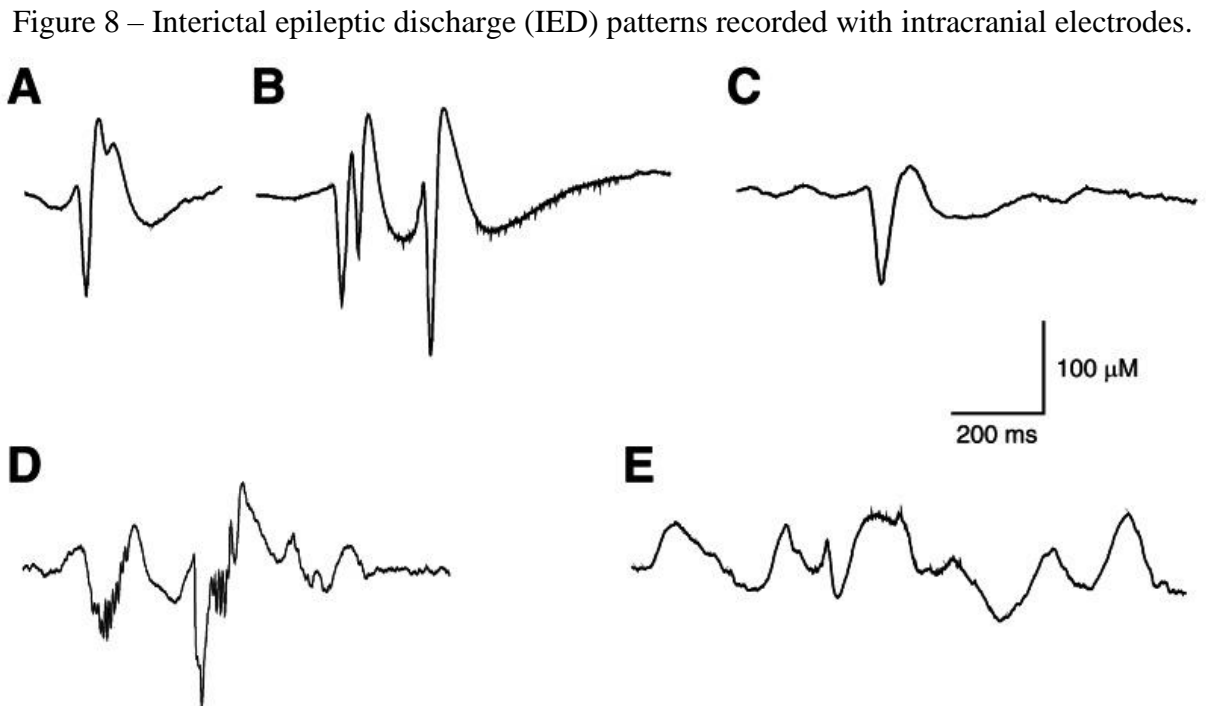
Figure 7 – IED Characteristics.



Source: From Bagheri et al. (2017)

Interictal discharges show a large pattern variability and may be divided morphologically into sharp waves, spikes, spike-wave complexes, polyspike-wave complexes, and sequences of fast oscillation. This diverse morphology may represent that they are generated by different neurobiological mechanisms and play different roles in the seizure generation. They are characterized by a large-amplitude rapid transient lasting 50–100 ms and are usually followed by a slow wave, 200–500 ms in duration, as shown in Figure 7. IEDs may

occur in isolation or brief bursts. Figure 8 shows several distinct possible patterns for the interictal discharges (IEDs).



Source: From Curtis et al. (2012)

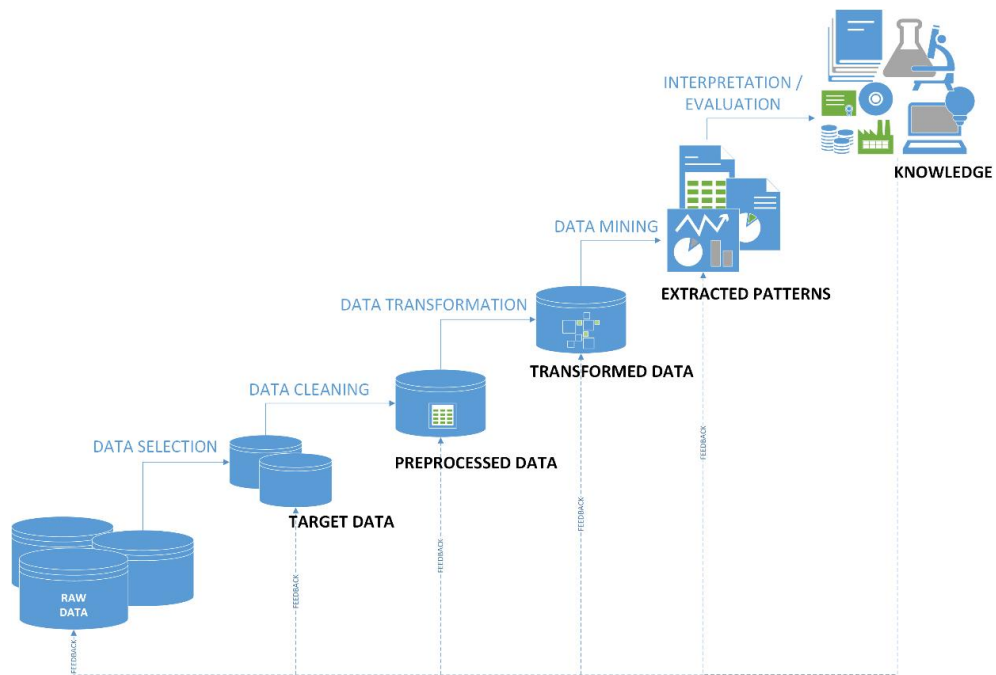
## 2.4 DATA MINING

Data mining (DM) is considered a fundamental part of a more extensive process called Knowledge Discovery in Databases (KDD) currently also called Data Science. Data Mining relates to the methods (algorithms, statistical tools, visualization techniques) to extract implicit, previously unknown, and potentially useful knowledge from data according to specifications or strategies, and to present them in an accessible and understandable form. Therefore, it is heavily dependent, for its correct application, on the previous steps of KDD. KDD and DM are also closely related to the fields of machine learning, statistics, pattern recognition, artificial intelligence, data visualization, among others (PACKT, 2016).

The KDD process is iterative and can contain loops between any two steps. It consists of the selection, cleaning, and transformation of data not only from databases but also from heterogeneous sources, applying to it data mining algorithms in order to discover valid, novel, potentially useful, and understandable hidden patterns (FUNES; DASSO, 2018).

It is possible to outline the basic steps of KDD as the following (Figure 9) (FAYYAD et al., 1996) (PACKT, 2016):

Figure 9 – Steps of KDD.



Source: Adapted from Fayyad et al. (1996).



1. **Understanding** – Developing an understanding of the application domain and the prior knowledge that is relevant to the defined goals.
2. **Data Selection** – Selecting the data from the available dataset, focusing on the subset of variables or samples that are appropriate for the analysis.
3. **Data Cleaning** – Applying basic techniques to remove noise, defining strategies to handle missing fields or changes.
4. **Data Transformation** – Defining useful features and fields that represent the data — applying dimensionality reduction or transformation.
5. **Data Mining** – Finding the appropriate data-mining method that matches the goals — searching for patterns using methods such as classification, regression, and clustering.
6. **Interpretation** – Understanding and interpreting the mined patterns, this can also involve the use of visualization techniques use of statistics and models and acting on the discovered knowledge.

These steps are the basis of the standardized DM methodology perspectives used by the industry like the Cross-Industry Standard Process for Data Mining - CRISP-DM (CHAPMAN et al., 2000) and Sample, Explore, Modify, Model, Assess - SEMMA (SAS, 2017), they all follow these same base steps but with different subdivisions or names (AZEVEDO; SANTOS, 2008) (SHAFIQUE; QAISER, 2014).

As stated before, the data mining process relates directly to the methods of extracting information rather than the entire process of dealing with data and the results.

The algorithms to mine the data tend to have two distinct goals, description and prediction.

The description focuses on finding new human-interpretable patterns to describe the data based on the available data set; the goals are to gain a more in-depth and non-trivial understanding of the analyzed data. On the other hand, the prediction involves using variables or fields from a database to build a model to predict unknown future values; they can be used to perform classification, prediction, estimation, among others. (AZEVEDO, 2017).

Some of the common tasks of data mining are:

- **Classification** – Using a function that maps the data into different predefined classes. The target variable should be of the nominal type.
- **Regression** – Finding a function that predicts some future numerical values using independent variables.
- **Clustering** – Using algorithms to identify finite sets of categories to describe the data grouping elements with high similarity.

Data mining methods can also be separated into three categories, according to the learning type: unsupervised learning, supervised learning, and semisupervised learning methods (KIM; SUKCHOTRAT, 2012).

- **Unsupervised learning methods** – These methods depend solely on the input variables and do not take into account the output information. Unsupervised learning aims to extract implicit patterns and elicit the natural groupings without using any information from the outputs. Examples of this category include k-means and principal component analysis (PCA).
- **Supervised learning methods** – These methods analyze the data using both the input and output variables to create the models that classify or predict the output values of future observations. Examples of this category include regression methods, decision trees, Support vector machines (SVMs), and artificial neural networks (ANNs).
- **Semisupervised learning methods** – These methods use a mixture of both unsupervised and supervised methods to generate an appropriate classification or prediction model. Examples of this category include support vector data description (SVDD).

The next subsections will explore the main tools used either by the data mining algorithm and by the analysis and mining of its results.

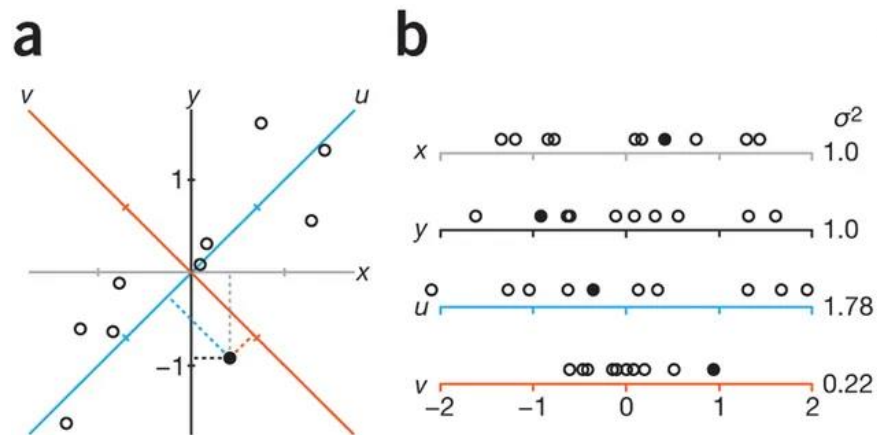
### 2.4.1 Principal Component Analysis - PCA

Principal component analysis (PCA) is a technique used for identifying principal components, which are a smaller set of uncorrelated variables, from a dataset that nonetheless preserves most of the sample's information. The technique emphasizes variation and captures strong patterns among the data, and it is used for dimensionality reduction. It is an unsupervised learning method similar to clustering, finding patterns without reference to prior knowledge about whether the samples come from different treatment groups or have phenotypic differences (LEVER et al., 2017). PCA is an important technique for a variety of fields, from image compression and face recognition to genetic analysis of populations (PENTAHO, 2019).

At first, it is essential to understand the concepts of eigenvectors and eigenvalues. An eigenvector is a direction of a new axis, the direction of the line can be vertical, horizontal, 45 degrees, and so on. An eigenvalue is a number, telling you how much variance there is in the data in that direction, telling us how spread out the data is on the new axis. The eigenvectors with the highest eigenvalues are, therefore, the principal components.

PCA is related to the covariance matrix of original variables, and the eigenvalues and eigenvectors are acquired from the covariance matrix. The product of the eigenvector corresponding to the largest eigenvalue and the source data matrix leads to the first principal component (PC), which expresses the maximum variance of the data set. The second PC is then obtained using the eigenvector corresponding to the second largest eigenvalue; this procedure is repeated  $n$  times to obtain  $n$  PCs, where  $n$  is the number of variables in the dataset. The PCs are uncorrelated to each other, and usually, the first few PCs are sufficient to account for most of the variations. Therefore, the PCA plot of observations using these first few PC axes facilitates the visualization of high-dimensional data sets (KIM; SUKCHOTRAT, 2012).

Figure 10 – PCA projections and PC calculations.



Source: From Lever et al. (2017)

Figure 10 represents 3 steps of PCA. In a) the axis  $v$  and  $u$  are defined. In b) all points are projected onto the lines, line  $u$  maximizes the variance ( $\sigma^2$ ) and it is, therefore, PC1. PC2 is the line perpendicular to PC1, in this case, line  $v$ .

### 2.4.2 $k$ -Means

The  $k$ -means algorithm is the most widely used clustering algorithm that uses an explicit distance measure to partition the data set into clusters. It is one of the unsupervised learning algorithms that solve clustering problems using a quantitative method. For a given a predefined number of clusters, it employs a simple algorithm to sort the data into groups.

The algorithm operates by doing several iterations of the same basic procedures.

It starts with initial estimates for the  $k$  centroids, which can either be randomly generated or selected from the data set. Each centroid is associated to one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance (TREVINO, 2016). All cluster centroids are then recalculated as the mean values of the instances that are assigned to specific clusters.

If the cluster assignments do not change at all, or if they have sufficiently few changes, the iterative process stops.

It is formally defined by the following objective function which minimizes the within-cluster sum of squares:

$$\operatorname{argmin} \sum_{j=1}^k \sum_{i=1}^n \|x_i - c_i\|^2$$

Where  $x_i$  represents a point on the data set of size  $n$ ,  $c_i$  is the centroid of cluster  $C_j$  and  $k$  is the number of clusters.

### 2.4.3 Maximal Likelihood Estimation

Maximum likelihood estimation involves computing the likelihood of the observed data as a function of the unknown parameters, based on the model to be fitted, and then determining the parameter values that maximize the likelihood.

The method of maximum likelihood is the most popular technique for deriving estimators. Considered for fixed  $\mathbf{x} = (x_1; x_2; \dots; x_n)$  as a function of  $\theta$ , the joint probability density (or probability)  $p_\theta(\mathbf{x}) = p_\theta(x_1; \dots; x_n)$  is called the likelihood of  $\theta$ , and the value  $\hat{\theta} = \hat{\theta}(X)$  of  $\theta$  that maximizes  $p_\theta(\mathbf{x})$  constitutes the maximum likelihood estimator (MLE) of  $\theta$ . The MLE of a function  $\tau(\theta)$  is defined to be  $\tau(\hat{\theta})$  (BESBEAS, 2012).

It solves the problem of modeling data distributions with several distributions to estimate the distribution that best describes the data. The actual formula of calculation depends on the selected probability function.

### 2.4.4 $p$ -Value

The  $p$ -values are calculated as a part of hypothesis testing, helping to define the statistical significance of the studied results.

Hypothesis testing is one of the most common methods for statistical inference. For a test of hypothesis, researchers define a hypothesis about population parameters and, based on extracted samples, verify its validity. The tested hypothesis is called the null hypothesis and is represented by  $H_0$ . The alternative hypotheses are commonly represented as  $H_a$  or  $H_1$ . The tested

hypothesis is always the null ( $H_0$ ) and not  $H_a$ . It is, therefore, a form of indirect proof, indicating its contradiction.

The testing process starts with the assumption that  $H_0$  is true for the population. If the magnitude of the difference between the obtained statistic and the population parameter is highly unlikely to be observed in a sample, then  $H_0$  is rejected in favor of  $H_a$ . If the observed difference is not sufficiently unlikely, then  $H_0$  is considered to be plausible for the population, so the rejection of  $H_0$  fails. Researchers either reject or fail to reject  $H_0$ ; The null hypothesis is never accepted because it is never proved, but rather the evidence may be insufficient to disprove it (CAPRARO; YETKINER, 2012).

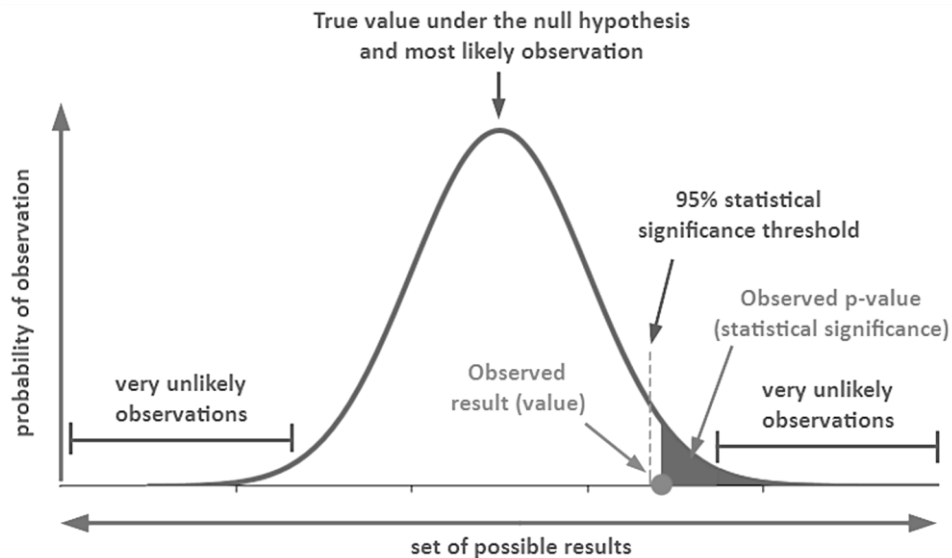
After the definition of the tested hypothesis and the alpha level ( $\alpha$ ), the  $p$ -value is calculated using the appropriate test, according to population characteristics, between one of the many statistical hypothesis tests available.

The alpha level is related to the defined research significance levels, and it is written in percentages. Typically used alpha levels are 0.05 (5%), 95% significance, and 0.01 (1%) with 99% significance (GLEN, 2014).

The interpretation guide to a given  $p$ -value reads as follows:

- A small  $p$ -value (typically  $\leq \alpha$ ) indicates strong evidence against the null hypothesis, so  $H_0$  is rejected.
- A large  $p$ -value ( $> \alpha$ ) indicates weak evidence against the null hypothesis, so  $H_0$  fails to be rejected.
- $p$ -values very close to the cutoff ( $\alpha$ ) are considered to be marginal. Therefore, nothing should be inferred.

One of the criticisms about  $p$ -values is that they are substantively affected by sample sizes. With large sample sizes, even very small differences or effects become statistically significant. By contrast, sometimes a statistical significance cannot be obtained because of the small sample size, although the effect might be present in the population (CAPRARO; YETKINER, 2012). Even though  $p$ -values can indicate which choice is more effective, it is fundamental that researchers provide an estimate of practical significance (i.e., effect size) when reporting  $p$ -values.

Figure 11 – Graphic representation of the  $p$ -Value.

Source: From Georgiev (2018)

Figure 11 shows that the  $p$ -value represents the area in the tail of a probability distribution, for this example the alpha was defined as 0.05.

#### 2.4.5 Wilcoxon rank-sum test

Wilcoxon rank-sum is a statistical hypothesis test that is often described as the non-parametric version of the two-sample  $t$ -test (WILD, 1997). Developed by Frank Wilcoxon in 1945, the test, instead of considering the direct values of the variables, replaces them with the rank scores. Since it is non-parametric, the test does not use group means and standard deviations as estimations of population parameters, therefore not assuming the normal distribution of the data (FAY; GEROW, 2013). It is applied when the populations might not be normally distributed, and the sample sizes of each group are small, on group sizes with fewer than 30 samples (FORD, 2017) (LAMORTE, 2017).

The test makes its inferences by determining the probability that two independently obtained groups are sampled from the same population. Considering two groups from a population that does not vary across the independent variable, there should be an equal probability that any score obtained will fall into either experimental group (PRATT, 2012). The

null hypothesis is, therefore, that data of two groups are samples from continuous distributions with equal medians, confronting the alternative that they are not.

In the process of calculation, the obtained values are ranked in order from smallest to largest, with the smallest value having rank 1, the 2<sup>nd</sup> smallest having rank 2, and so on. Given the hypothesis, each rank should have an equal chance of belonging to either one of the groups.

The primary strength of the Wilcoxon rank-sum test is that it does not require normal distribution for small sample sizes. The transformation of observations into rank scores attenuates the impact of outliers on the statistics. For normally distributed data, the Wilcoxon statistic shows the same power as the t-test.

The determination of the  $p$ -value for the Wilcoxon rank-sum statistic involves calculating the probability of obtaining a given rank-sum score by chance from the null population, out of all possible combinations of summed rank scores gathered from the two groups (PRATT, 2012). For larger sample sizes, the calculation is done using z-scores.

#### **2.4.6 Effect size statistics**

The effect size statistics quantify the relationship between groups. They are especially significant in the medical field since it explains how substantial an effect of the analyzed variable is.

An effect size is a specific numerical nonzero value that is used to represent the degree of difference between the two populations in those occurrences for which the null hypothesis was estimated false. In the cases in which the null hypothesis is false (rejected), the results of a test of statistical significance indicate that consistent differences exist between two populations on the phenomenon of interest, but test outcomes do not provide any value regarding the extent of that difference (PIASTA; JUSTICE, 2012).

#### **Cohen's $U_3$**

As an effect size, Cohen's indexes are typically used to represent the magnitude of differences between two (or more) groups on a given variable. When comparing means in a scientific study, the reporting of an effect size such as Cohen's are considered complementary



to the reporting of results from a test of statistical significance. The calculation of Cohen's indexes and their interpretation provides a way to estimate the actual size of observed differences between two groups, explicitly, whether the differences are small, medium, or large (PIASTA; JUSTICE, 2012).

There are several versions of Cohen's indexes. In this work, Cohen's  $U_3$  was chosen because it has been rated as the most informative one (HANEL; MEHLER, 2018).

It is calculated by the given formula:

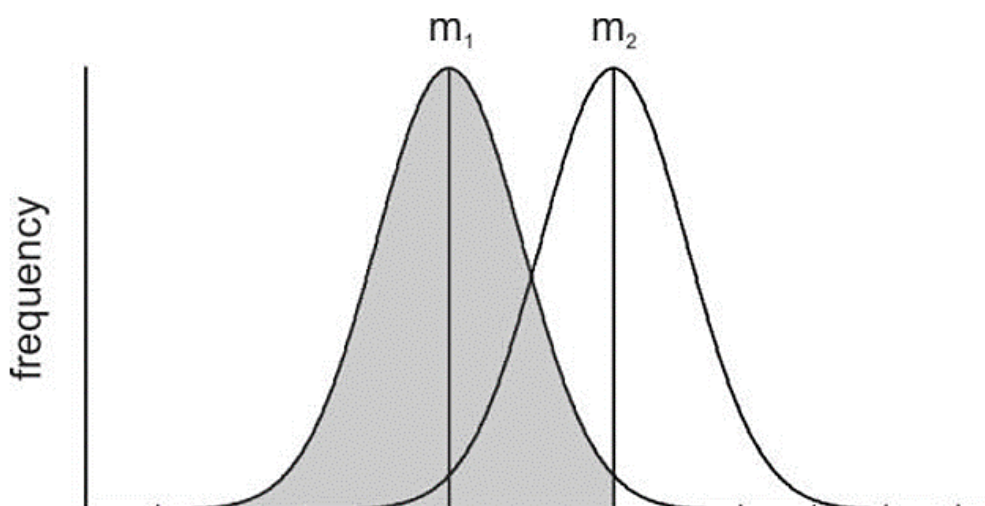
$$U_3 = \frac{n_{X < \text{median}(Y)} + 0.5 n_{X = \text{median}(Y)}}{n_X}$$

Where  $n_{X < \text{median}(Y)}$  is the number of elements in group X that are exceeded by the median value of group Y,  $n_{X = \text{median}(Y)}$  is the number of elements in group X that are equal to the median, and  $n_X$  is the total number of elements in group X.

Interpretation guide:

The ranges of Cohen's  $U_3$  vary from 0 to 1, with 0.5 describing "no effect". Cohen's  $U_3$  gives the proportion of scores in one group that are smaller than the typical value (i.e., the median) of the other group.

Figure 12 – Illustration of Cohen's  $U_3$



Source: From Hentschke (2018)

Figure 12 displays an example of the interpretation of Cohen's  $U_3$ . Two standard normal distributions are shown. The gray shaded area marks the proportion of distribution 1 that is exceeded by the median of distribution 2.

### Hedges' $g$

Hedges'  $g$  is another measure of effect size. It usually expresses the difference between an experimental group and a control group. It is similar to Cohen's  $d$ , except that it outperforms Cohen's  $d$  in smaller sample sizes ( $< 20$ ) (GLEN, 2016).

It is calculated by the given formula:

$$g = \frac{m_1 - m_2}{S_{within}}$$

Where  $m_1$  is the mean of the first group,  $m_2$  is the mean of the second group, and  $S_{within}$  is the pooled standard deviation, which is the square root of the pooled within-groups variance, weighted by the degrees of freedom in each group, as seen below:

$$S_{within} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Hedges'  $g$  is known to show biased results, about 4% overestimate, for small group samples ( $< 50$ ) (GLEN, 2016). Hedges (1981) described an approximate bias correction formula:

$$g_{unbiased} = g_{biased}c(df_{within})$$

Where:

$$c(df_{within}) = \left[ 1 - \frac{3}{4df_{within} - 1} \right]$$

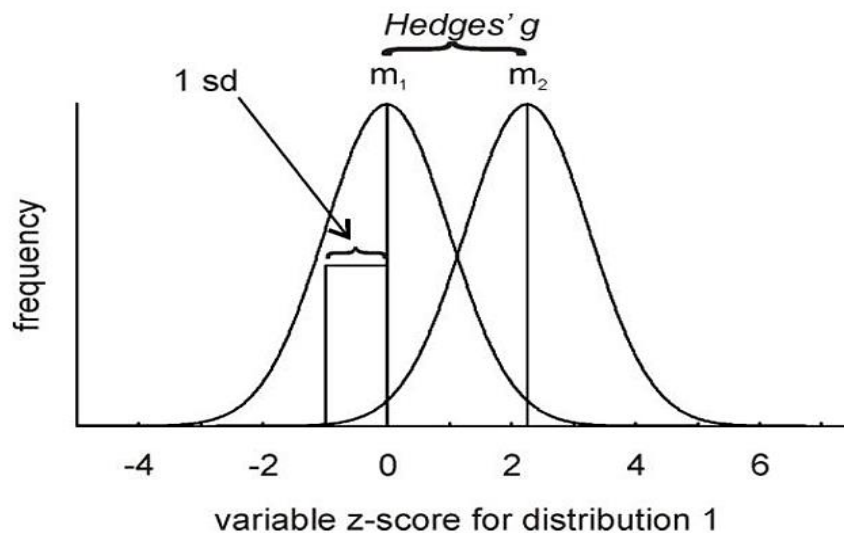
And  $df_{\text{within}}$  are the degrees of freedom used to compute  $s_{\text{within}}$ , namely  $n_1 + n_2 - 2$ , and  $n_1$  and  $n_2$  are the sizes of each group.

Interpretation guide:

The  $g$  value indicates how many standard deviations one group differ from another. Standard deviations are equivalent to  $z$ -scores (1 standard deviation = 1  $z$ -score) (HENTSCHKE, 2018).

- Small effect (cannot be perceived by just looking at the data)  $\approx 0.2$
- Medium Effect  $\approx 0.5$
- Large Effect (can be perceived by just looking at the data)  $\approx 0.8$

Figure 13 – Illustration of Hedges'  $g$ .



Source: From Hentschke (2018)

Figure 13 illustrates the meaning of Hedges'  $g$ . Two standard normal distributions are shown, with means  $m_1$  of 0 and  $m_2$  of 2.2, and a standard deviation (SD) of 1 (identical for both distributions). According to this example, Hedges'  $g = (2.2 - 0) / 1 = 2.2$ .

## Omega Squared ( $\omega^2$ )

Omega squared ( $\omega^2$ ) is a descriptive statistic used to measure the strength of the relationship between a qualitative variable and a quantitative variable, estimating how much variance in the quantitative variables are accounted for by the qualitative variables. It complements the results of hypothesis tests comparing two or more population means (OLEJNIK, 2012). It is viewed as a lesser biased alternative to eta-squared when given small sample sizes.

It is calculated by the given formula:

$$\omega^2 = \frac{SS_{effect} - (J - 1) MS_{error}}{SS_{total} + MS_{error}}$$

Where  $SS_{effect}$  is the sum of squares between groups,  $SS_{total}$  is the overall sum of squares,  $J$  is the number of levels of the factor (groups), and  $MS_{error}$  is the mean squared error within groups (HENTSCHKE, 2018).

Interpretation guide:

Omega squared ( $\omega^2$ ) value clarifies how much variance in the metric variable is explained by group membership. It varies between 0 and +1; with 0 meaning “no effect”.

### 2.4.7 Classification Models

#### ***k*-nearest neighbors (*k*NN)**

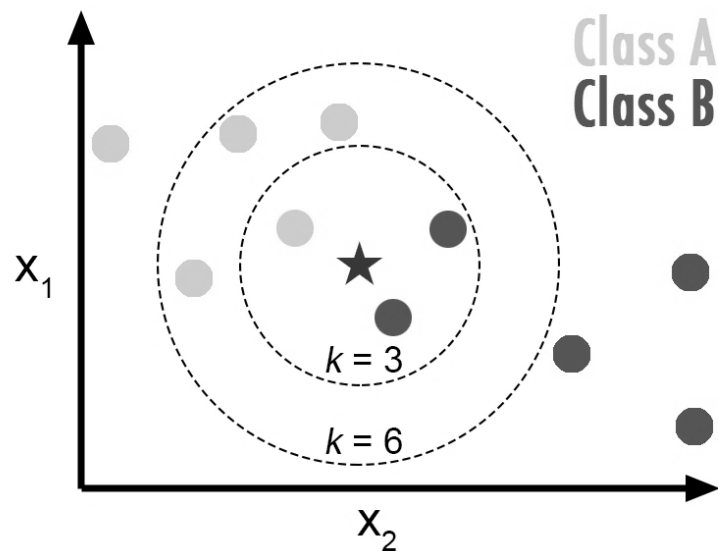
The *k*-nearest neighbors (*k*NN) is a non-parametric, lazy-learning (instance-based learning) technique, and it is one of the simplest and most commonly used learning algorithms. It is employed for credit ratings, political science, image and video recognition, among others. *k*NNs do not have an explicit training phase; it collects data from a training data set and uses this data later to make predictions for new records. The algorithm is based on feature similarity,

using a majority voting mechanism (IBM, 2018); a point is classified by a majority vote of its neighbors, therefore being assigned to the most common class among its  $k$  nearest neighbors.

This technique is also useful for regression (prediction of values); in this case, the new value is the average (or median) of the values of its  $k$  nearest neighbors.

The algorithm works as the following, given a query point, the  $k$  closest points are determined. A variety of distance measures can be applied to determine how close each point is to the checked point. Then, the  $k$  nearest points are analyzed to find which of the categories they belong to. Finally, this category is assigned to the checked point of the time. This procedure repeats for all points that require classification (KIM; SUKCHOTRAT, 2012).

Figure 14 – Example of  $k$ NN classification.



Source: From ACM (2016)

Figure 14 represents an example of  $k$ NN classification. The test point (star) should be classified either to the class B or to the class A. If  $k = 3$  (inner circle), it is assigned to the class B because there are 2 class B and only 1 class A inside the inner circle. If  $k = 6$ , it is assigned to the class A (4 class A vs. 2 class B in the outer circle).

## **Ensemble**

Ensemble learning model is a well-established set of machine learning and statistical techniques for improving predictive performance through the combination of different learning algorithms. The combination of the predictions from different models can generally increase the accuracy strengthening the performance of the ensemble model. Ensemble methods come in different flavours and levels of complexity, depending on the combination of multiple deep learning networks (PINGEL, 2019). Additional applications of ensemble include assigning a confidence to the decision made by the model, selecting optimal (or near optimal) features, data fusion, incremental learning, nonstationary learning and error-correcting (ILIADIS; JAYNE, 2015).

The functioning of model can be based on averages, weighted averages or voting process, and varies greatly depending on the combination of the support models.

Although not so popular in the deep learning literature as it is for more traditional machine learning research, the ensemble models produce impressive results, which can be attested by the winning of popular machine learning competitions, such as ImageNet and Kaggle challenges.

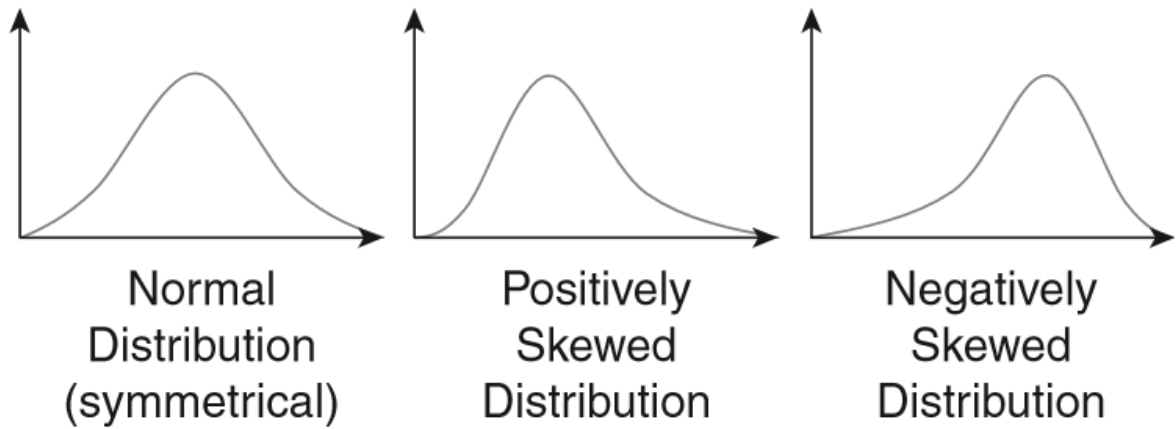
### **2.4.8 Descriptive Statistics**

Descriptive statistics include measures of central tendency, dispersion, shape, correlation, and covariance of giving data. Since some of them are well known and trivial (i.e. mean, median, standard deviation...), the focus will be to offer a brief introduction to some more unfamiliar statistics.

#### **Skewness**

Skewness is a measure of symmetry, more precisely, the lack of symmetry. Negative values of skewness indicate that the data is skewed left, and positive values for the skewness indicate that data is skewed right (NIST/SEMATECH, 2012).

Figure 15 – Example of distributions with different skewness.



Source: From Mihaescu (2012)

## **Kurtosis**

Kurtosis is a statistical measure of shape that defines how strongly the extremes of a distribution differ relative to a normal distribution. A set with high kurtosis represents that the extremes of a given distribution contain extreme values (NIST/SEMATECH, 2012). Along with skewness, kurtosis is an essential descriptive statistic of data distribution.

### **2.4.9 Statistical Visualization**

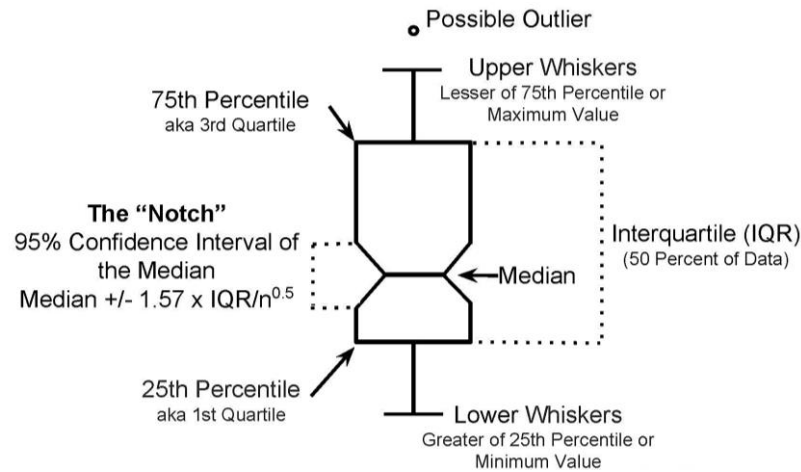
Statistical visualization is the use of graphs, plots, and infographics to represent data clearly and efficiently. Many types of graphs are well known such as histograms and scatterplots. Below there is a brief description of a specific version of the box plot diagram as well as some graphs used to display the performance of prediction models.

#### **Notched Box Plot**

A Notched Box Plots are a handy graphic way to display many relevant characteristics of the data. It displays the interquartile ranges, turning it possible to visually inspect the outliers,

the median, and also contains a notch that facilitates comparison between two groups, showing with 95% confidence if their medians differ (DOYLE, 2016).

Figure 16 – Notched box plot characteristics.



Source: From Doyle (2016)

## Confusion Matrix

It is a performance measurement for machine learning classification problem where output can be two or more classes. For two classes (i.e. 1 and 2), it is represented as a table with 4 different combinations of predicted and actual values (Figure 17), the values can show the number or the percentages of predictions. Measures of accuracy, recall, and precision can be calculated using a confusion matrix (NARKHEDE, 2018).

Figure 17 – Example of a confusion matrix for two classes.

	<i>Class 1 Predicted</i>	<i>Class 2 Predicted</i>
<b>Class 1 Actual</b>	TP	FN
<b>Class 2 Actual</b>	FP	TN

Source: From Sharma (2019)



Each position of the matrix represents one of the scenarios of prediction. For the example of classes 1 and 2, the interpretation of each position is described below:

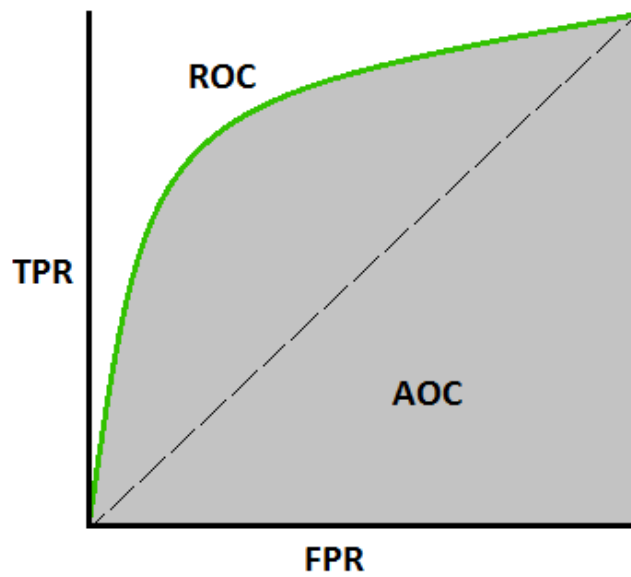
- True Positive (TP): Predicted class 1 and it is right;
- True Negative (TN): Predicted class 2 and it is right;
- False Positive (FP) - Type 1 Error: Predicted class 1 and it is wrong;
- False Negative (FN) - Type 2 Error: Predicted class 2 and it is wrong.

### **ROC Curve**

The ROC (Receiver Operating Characteristics) curve is a performance measurement for classification problems at various threshold settings. It is one of the most important evaluation metrics for checking any classification model's performance and it is usually used in conjunct with the area under curve (AUC) (NARKHEDE, 2018). The ROC is a probability curve and the AUC represents a degree or measure of separability. It tells how much model is capable of distinguishing between classes. The AUC index varies between 0 and 1, where a higher value means a better prediction model.

The ROC curve is plotted with the true positive rate (TPR) against the false positive rate (FPR) where TPR is on y-axis and FPR is on the x-axis as seen in Figure 18.

Figure 18 – Representation of a ROC curve.

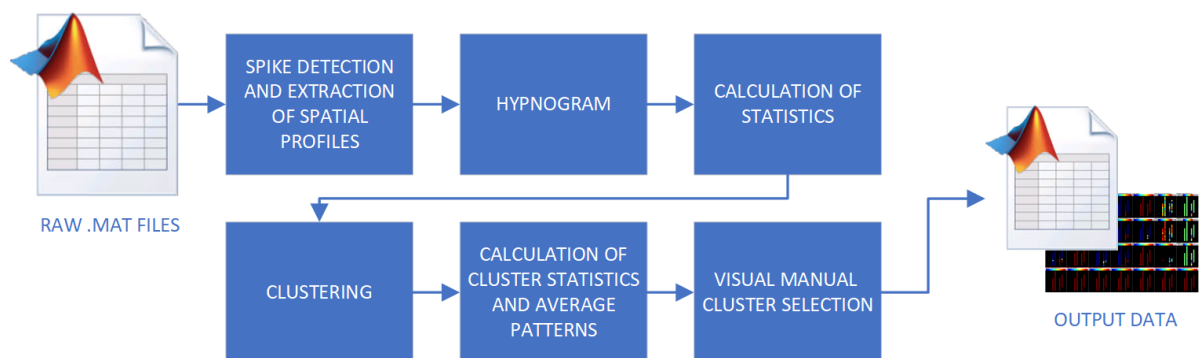


Source: From Narkhede (2018)

## 2.5 EEG ANALYSIS TECHNIQUE

The chosen EEG data mining algorithm is composed of six parts: spike (IEDs) detection and extraction of spatial profiles, hypnogram, calculation of overall statistics, clustering, calculation of clustering statistics and average patterns, and cluster selection (Figure 19). The inputs of the algorithm are the raw interictal EEG recording data, stored in MATLAB<sup>®</sup> matrices after the selection and cleaning process. The outputs given are the cluster statistics, overall EEG statistics, and visual representation of the detected cluster activity.

Figure 19 – Scheme of the data mining algorithm.



Source: From the author.

The spike detection is done by a robust detection algorithm that adaptively models statistical distributions of signal envelopes allowing a reliable distinction between the IEDs and the background data (Figure 20). The detection captures even low-amplitude IEDs that are often subject to oversight. It also proved to have a better performance when compared to both human readers and reputable detection software (JANCA et al., 2015).

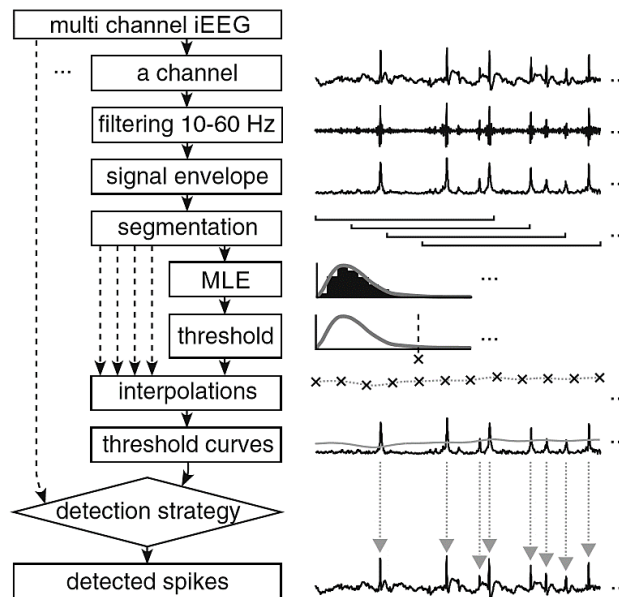
Since spikes are characterized by a large-amplitude rapid transient lasting 20–70 ms, they are, in the frequency spectrum, characterized by a local increase, particularly in the 10–60 Hz frequency bands (given that  $f = T^{-1}$ ). Signals with higher frequency are resampled to 200 Hz to maintain filter characteristics. They are then subjected to the filtering process, a 10–60 Hz pass-band, and another 4 Hz band in order to reduce the noise.

An instant envelope of each filtered channel is calculated using the absolute value of the Hilbert transform (WITTE et al., 1991). Spikes induce an increase in energy, which manifests as a peak in the envelope. The algorithm estimates the statistical distribution of the envelope and identifies a threshold value, which enables discrimination of spikes from background activity (JANCA et al., 2015).

IEDs are detected sequentially through channels. A detection within 5 ms difference from the previous IED detection in a different channel is considered a multichannel IED event (JANCA et al., 2018).

The statistical distribution of the envelope is calculated for each segment and approximated with the best-fitting statistical model using a maximal likelihood algorithm (MLE). A log-normal model is utilized since it provides the best description of the data characteristics. The maximal amplitude in each channel is determined from the signal envelopes (JANCA et al., 2013). The spatial profile of the amplitudes within each IED display the relevant information on how they spread across the brain.

Figure 20 – Spike detection algorithm scheme.



Source: From Janca et al. (2015)

The data of the IED events are stored in matrices. Each column of the matrix represents one event in time, and rows represent bipolar channels.

A matrix  $Q$  stores the binary information about a detected IED: 1 – detected IED, 0 – no detection. A Matrix  $S$  with the same dimensions stores the values of the maximal amplitude of the envelope through the event (JANCA et al., 2013).

For the extraction of the spatial profiles of the events, the matrices  $Q$  and  $S$  had to be pre-processed in order to reduce dimensionality and to enhance the stability of the sorting procedure. To remove high amplitude artifacts, Tukey's rule (SULLIVAN; LAMORTE, 2016) is used discarding the outliers from matrix  $S$ .  $k$ -means is applied to the IED rates to discard false positives and reduce the dimensionality on both matrices.

The PCA technique is then applied to the pre-processed, non-centered matrix  $S$  to extract frequently occurring spatial profiles of IED events by the calculation of the principal components.

For real EEG data, a threshold criterion is used to identify and to extract only significant (high eigenvalue) components. The threshold is defined after randomizing the matrix  $S$  100 times and processing it by PCA. The threshold is composed by the 95% percentile of the largest eigenvalues of the randomized matrix  $S$  (JANCA et al., 2015).

After the detection of IEDs and the definition of their spatial profiles, there is an intermediate section called hypnogram. This part separates and labels the data into the two vigilance periods, awake and sleep. This separation is necessary given the diverse epileptical activity behavior in these two stages, as discussed in subsection 2.2.1. The selection is made manually, using MATLAB's graph tools. The awake patient activity usually consists of morning and afternoon recordings. For the sleep data, the detected activities were recorded between midnight and 3 AM.

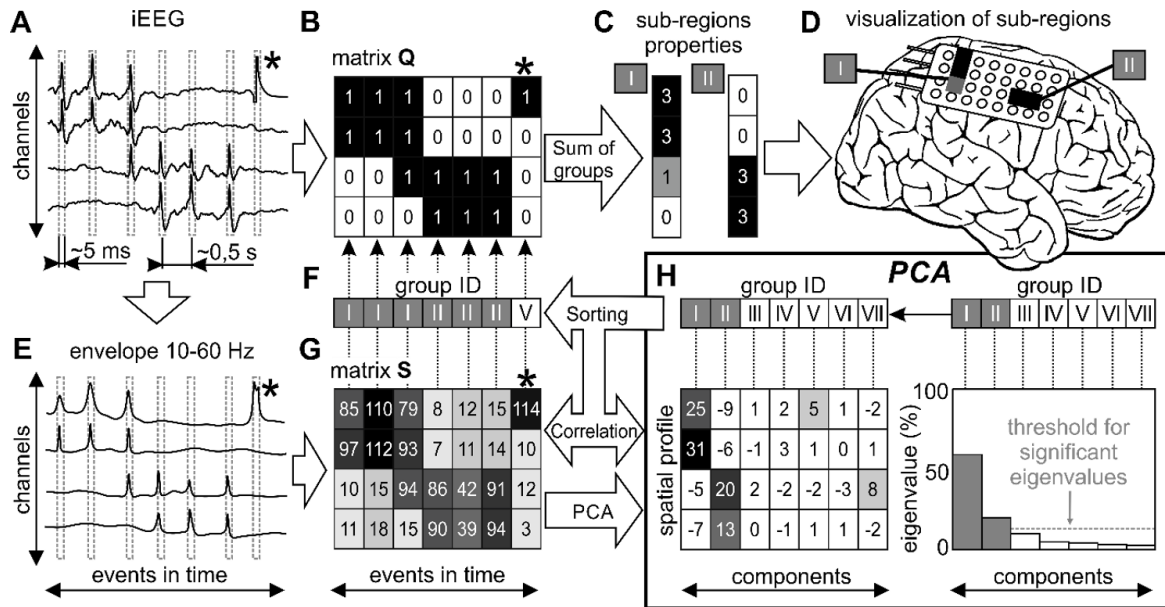
The third part of the algorithm is the calculation of overall statistics. That means the calculation of important metrics about the detected activity, such as the number of detected IEDs, and the IED rate (IEDs/min).

The fourth step is the clustering process. The clustering algorithm uses PCA to identify patterns of how the IED spread within the brain; the calculated spatial profiles are assigned to a principal component using Pearson's correlation (SIGMA PLUS STATISTIEK, 2019).

Each detected event (columns of the S matrix) is correlated with all components and assigned through minimal correlation distance to the best fitting spatial profile. Since PCA transformations can be rotated, it can generate positive and negative components. Therefore, each event can be correlated with both the positive and negative versions of the components, and the higher correlation result is accepted. The sorting procedure results in groups of IED events with similar spatial profiles. If two groups share a spatial profile with a correlation higher than 90%, they are merged into a single group (JANCA et al., 2018).

The algorithm also supports a coefficient setting for a low (0) or high (1) separability approach for the clustering process; the coefficient affects the merging process of the groups with a close spatial profile. Previous tests report that the high separability coefficient produces an output with considerable higher statistical significance (INÁCIO; JANCA, 2019).

Figure 21 – IED sorting according to their spatial profile (clustering process).



Source: From Janca et al. – Supplementary material (2018)

After the clusterization process, represented in Figure 21, where the sub-regions correspond to the clusters, the algorithm calculates new statistics such as total activity detected per cluster and the percentage of the total activities for each cluster. It also calculates the average wave patterns displayed on each sub-region generated from raw iEEG data, preparing it for the next step.

The last step is the visual inspection and selection of clusters. The average waveforms for each cluster are plotted using MATLAB graphing tools. Each cluster is then visually inspected in order to discard any artifacts or false positives that were still present among the analyzed data. The post-processing of the sorted events generates a quantitative description of the functional organization of the epileptical networks in the form of clustered activity data. The pos-processed data, including previous calculation of statistics, are saved into MATLAB matrices. The graphical representation of the selected sub-regions is also created (JANCA et al., 2018). These represent the actual outputs of the algorithm.

## 3 METHODS

### 3.1 PRE-PROCESSING

Even though this work focuses on the DM application, it is important to describe some of the previous steps that can be identified as part of KDD, they were, with the exception to the last, all done by third parties. These intermediate stages are needed since they serve as preparation for the application of algorithms and to make sure that the data is ready for the extraction of the information.

Before starting the analysis, it is also essential to understand the data and the application domain, bring all the prior knowledge necessary up, and to set the goals. These were either done by the exploratory research (bibliographical), the internship experience and were detailed on the previous chapters of this work.

The first part of the long path of the data analysis begins with the source of raw data. The data was recorded in Prague, Czech Republic, at the Motol University Hospital of Charles University in a partnership with the Czech Technical University in Prague (CTU). Data collection was approved by the institutional ethics committee, and official personal or parental consent was obtained. Dataset consists of long term iEEG recordings, clinical evaluation of brain epileptogenic zones, and the clinical definition of after-surgery outcome. The datasets are made available for studies of the Intracranial Signal Analysis Research Group - ISARG due to the partnership.

The raw data of the iEEG recordings from each patient go through a first procedure, to turn the readings into readable MATLAB<sup>®</sup> matrix files, and to separate the data into ictal and interictal, using annotations from the video iEEG monitoring made available from the hospital or seizure reporting by clinicians. This treatment of the data can be defined as the first data cleaning process these recordings are subject to since it eliminates noise and all irrelevant sections of the recordings. The files are, at that point, labeled with patient numbers and time and date of the records and stored in a server at CTU.

With all the available information, the files are ready for the data selection, which is the next step of the KDD process. All the files that are relevant for this work were selected and retrieved from the data collection server. In this case, the selection consists of awake and sleep

vigilance epochs of the interictal recordings of FCD epilepsy patients with available postoperative outcome information. These parameters reduced the scope of the analysis from about a hundred to 46 patients that have met all the criteria. Another group of 6 patients was also selected and put on a waiting list since they only lack surgery outcome data.

### 3.2 DATA AND PATIENTS

In this section, some characteristics of the data and the patients will be described.

The MATLAB<sup>®</sup> files with the raw iEEG recordings from 52 patients take 120 GB of hard disk space, and after the data mining process, it increases up to 200 GB.

Each patient also has a spreadsheet file with medical information provided by the hospital. Due to restrictions on the availability of some variables, the sample sizes for each analysis differ from the total number of patients.

Table 5 contains some patient's information extracted from the spreadsheets provided by the hospital.

Table 5 – Patient Characteristics.

Patient Number	Gender	Age at Surgery	Years of Epilepsy	FCD type	Follow-Up (Years)	Engel
P005	M	17	11	2A	7	III
P012	F	37	23	2B	10	IIIa
P025	M	4	4	2B	7	I
P030	F	17	14	2B	7	III
P033	F	16	6	1B	6	I
P034	F	9	7	1	8	Ia
P035	M	41	22	2B	7	I
P036	F	44	26	2B	8	I
P038	M	32	0	2B	8	I
P043	F	6	6	2B	2	Ia
P045	F	9	7	1	6	I
P046	M	7	7	2A	6	I
P048	M	45	18	3(1A)	6	IV
P060	M	54	28	1B	10	Ia
P063	F	41	23	1B	9	Ia
P066	F	35	22	2B	5	I
P067	M	12	11	2B	5	IV
P068	M	33	18	1A	2	IVb
P072	M	17	8	1	5	IV



P074	M	34	24	2A	4	IIa
P075	M	16	10	2B	1	III*
P078	M	34	24	1A	3	IVb
P079	F	33	18	1A	4	Ia
P082	F	33	15	2B	3	Ia
P084	M	4	3	1B	4	Ia
P085	F	37	32	3(2B)	9	Ia
P091	F	3	2	2B	3	Ia
P096	F	30	18	3(1B)	3	Ia
P097	M	30	29	1B	3	Ia
P110	F	23	4	1B	3	Ia
P117	M	9	9	2B	2	Ia
P119	M	37	32	2B	3	Ia
P125	M	15	11	2B	1	III
P126	M	19	10	2B	3	Ia
P127	M	8	2	1	2	I
P129	F	30	10	3(1A)	2	IV
P133	F	10	5	IA	2	I
P136	M	33	24	3(1B)	1	IVb*
P142	M	28	8	3(1A)	2	Ia
P144	M	54	44	2B	2	Ib
P147	F	29	29	2B	2	Ia
P155	M	24	9	1A	1	Ib*
P162	F	43	28	2B	1	Ic*
P165	M	34	33	2B	1	Ia*
P170	M	48	28	1	1	IIIa*
P173	F	38	37	3(1A)	1	IVa*
P143	M	14	3	1A	1	-
P150	F	12	2	2B	1	-
P163	M	2	2	2B	1	-
P177	M	18	16	1A	0	-
P179	F	18	2	1A	0	-
P185	F	33	33	2B	0	-

Legend: Yellow lines - Surgery outcome still unavailable. \* - Preliminary outcome

Source: From the author.

Table 6 details the sample sizes for each analysis in this work.

Table 6 – Number of patients in each analysis.

<b>Statistical Analysis</b>	
Sleep + Awake	41
Sleep	42
Awake	45
Untested (lack of outcome)	6
<b>Predictive Analysis</b>	
Training Group	28
Test Group	11
Untested (FCD III)	7

Source: From the author.

### 3.3 DATA MINING

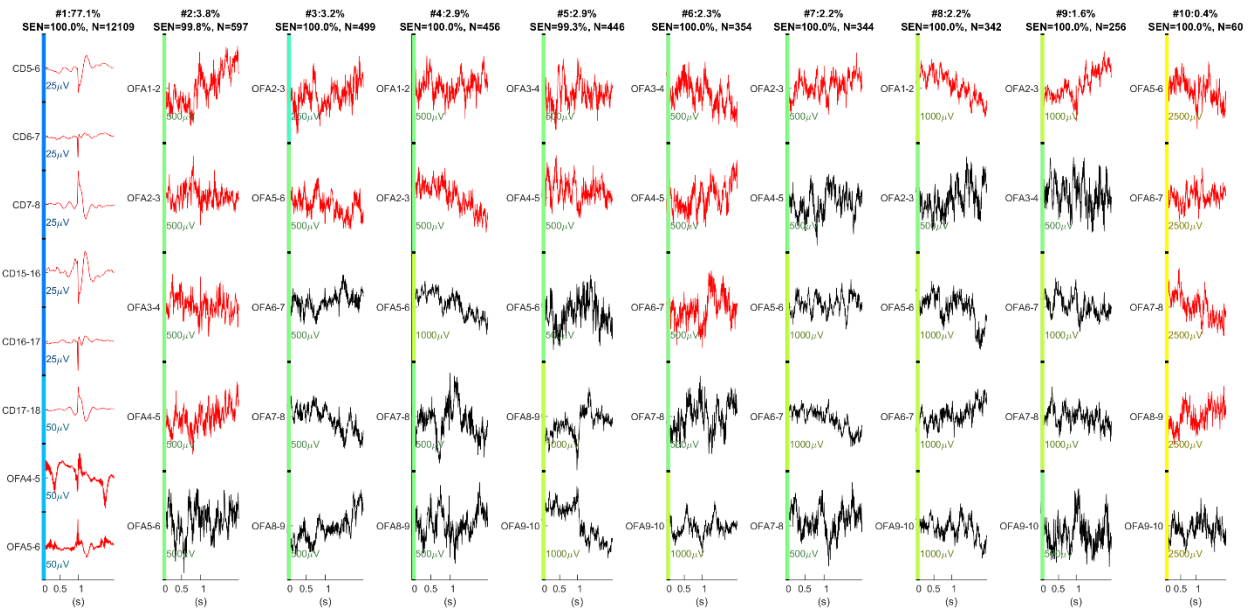
The chosen algorithm performs multiple tasks and can be described as a detection, analysis and clustering data mining algorithm. It detects and groups the IEDs according to the onset and spreading characteristics and patterns. All the details about the selected technique are were discussed in subsection 2.5 of Chapter 2.

The execution of the algorithm is straightforward. The inputs are the iEEG matrices. Additional settings are necessary to set the separability coefficient (1 – High separability) and for the removal of error channels, available in the patient's spreadsheets from the hospital.

One section of the execution that deserves some attention is the manual selection of the clusters. Given the characteristics of the artifacts, some may not be eliminated through the detection process, as seen by the false positive rate of the algorithm. Discarding clusters that do not display epileptical activities is an important step for the proper analysis of the relevance of the algorithm's outputs.

The methodology for picking the clusters is to choose the maximal number of valid clusters until reaching clusters with around 0.3% of detected activities, or until the very last, if the total number of clusters is ten or less. Very small clusters have minimal impact on the calculated statistics and would only increase data dimensionality.

Figure 22 – Example of cluster selection.



Source: From the author.

Figure 22 shows the selecting interface for a patient. Each column represents a cluster and each line represents a channel with its respective name on the left side. At the top, beside the cluster number, it shows the contribution of each cluster for the total detected activity as percentage. “N” displays the number of detected IEDs for a cluster. Red lines indicate that the detected patterns are among the 5% most active, black ones are in the active region but below 95% percentile.

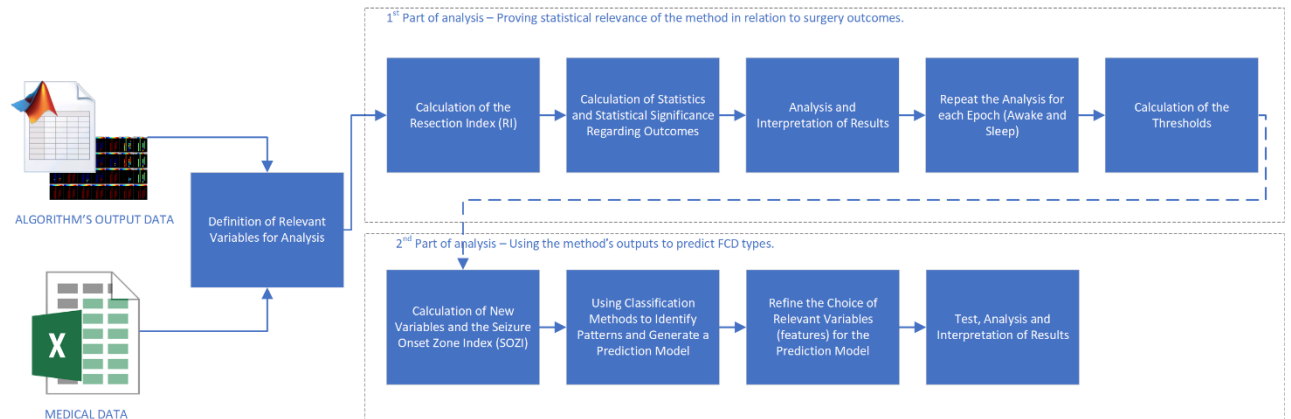
Cluster number #1 (first column), with 77.1% of the detected activity, is the only that shows the typical spikes that characterize the interictal epileptiform discharges (IEDs). All other clusters display signal waves that are typical of artifacts and should be discarded.

After the selection of clusters, the run of the algorithm is over and all the detected IED data, the statistics, and the graphs are saved. The outputs of the algorithm, along with the medical information of the patients, are the basis for the statistical analysis that composes the first investigation.

### 3.4 INTERPRETATION AND EVALUATION

Figure 23 shows the steps of interpretation and evaluation that were applied to the data after the data mining step.

Figure 23 – Interpretation and evaluation steps.



Source: From the author.

The algorithm offers a vast number of variables and measures that describe features mined from the recordings, most of them related to the detected IEDs.

Table 7 describes the selected variables that are relevant for achieving the analysis goals.

Table 7 – Description of the relevant output variables.

Output Variable	Description
<b>Number of Clusters</b>	Number of detected clusters after manual selection.
<b>Cluster Total Activity</b>	Total number of detected IEDs in each cluster.
<b>Cluster Activity as % of Total</b>	Percentage of the activity present in each cluster.
<b>Number of Active Region (AR) Channels</b>	Number of channels in each cluster with higher activity (active) computed using the k-means method.
<b>Channel Activity in each Cluster</b>	Rate of IED in each channel, by cluster.
<b>Detected IED Rate (qEEG)</b>	Rate of detected IEDs weighted by a factor number (ambiguous = 0.5, obvious = 1).

Source: From the author.

Each patient's spreadsheet also contains a significant amount of information regarding the diagnostic iEEG recordings. Table 8 contains a description of the selected variables that are relevant to the analysis.

Table 8 – Description of the relevant medical variables.

<b>Medical Variable</b>	<b>Description</b>
<b>Surgery Outcome (ENGEL)</b>	The clinical definition of the patient's outcome after the surgery.
<b>Clinical Evaluation Data of SOZ</b>	The clinical definition of the brain's area (electrode placement) classified as the seizure onset zone (SOZ).
<b>Surgery Resection Area</b>	Clinical information on the resected brain areas in the surgery.
<b>Patient's FCD Type</b>	Clinical definition of the Focal Cortical Dysplasia (FCD) types.

Source: From the author.

Not all data is ready for the analysis by default; in most cases, some treatments are needed prior to the calculation of statistics and indexes.

### 3.4.1 Treatment of Zeros

In the patient's medical spreadsheet from the hospital, some channels are marked as containing errors, usually technical artifacts. These channels have their values zeroed during the analysis and exportation of the data and are ignored during the calculation of indexes and statistics. Electrocardiogram (EKG) channels must be removed since they are employed to detect biological artifacts that contaminate EEG measures. For some patients, the removal of this channel was not indicated on the spreadsheet, probably due to overlooking, this was corrected before running the algorithm.

### 3.4.2 Calculation of Outliers

The false-positive rate of the IED detector is  $2.4 \pm 2.4$  IEDs/min (JANCA et al., 2015). This can suggest that patients with low detected IED activity may display a large portion of false positives among the detected activity. Using a strict rule of mean plus two standard

deviations, the base value of detected activity for outliers would be 7.2 IEDs/min. Any patient with max qEEG rate less this value was excluded from the analysis. The max qEEG rate for each patient was chosen instead of the average rate, given that patients with very focalized activities may display a low average, but a high rate of detection on the electrodes positioned directly on the epileptogenic areas. Despite that, all statistics were calculated with and without the outliers to measure their impact on the results.

### 3.4.3 Calculation of the Resection Index - RI

In possession of all information, it is possible to calculate the resection index (RI). The index provides information about the overall IEDs detected by the channels, weighted by cluster activity, included in the resection area (JANCA et al., 2018).

The choice of the RI to verify a correlation with the surgery outcomes, comes from the fact that, given the difficulties to accurately delineate the epileptogenic zone (EZ), one can consider that in surgeries with a good outcome, the epileptogenic zone was resected, and on the contrary, for a poor outcome the EZ was not completely resected. The RI is defined by the formula below:

$$\mathbf{RI} = \sum_{cl} W_{cl} A_{cl}$$

Where  $A_{cl}$  is the weight of a cluster, the number of detected activities in a cluster as a percentage of the total detected activity. And  $W_{cl}$  is defined as:

$$W_{cl} = \frac{\sum_{ch \in RES} Q_{ch}}{\sum_{ch} Q_{ch}}$$

$W_{cl}$  represents the sum of the detected activities (Q) of each channel inside the resected area (RES), divided by the sum of detected activities of all channels, for a cluster.

The index must be calculated for each considered epoch. For this work, three scenarios are considered. That is, the awake and sleep activities analyzed together, and awake and sleep separately.

Except for some patients that do not have recordings for one of the epochs, patients have three resection indexes (RI).

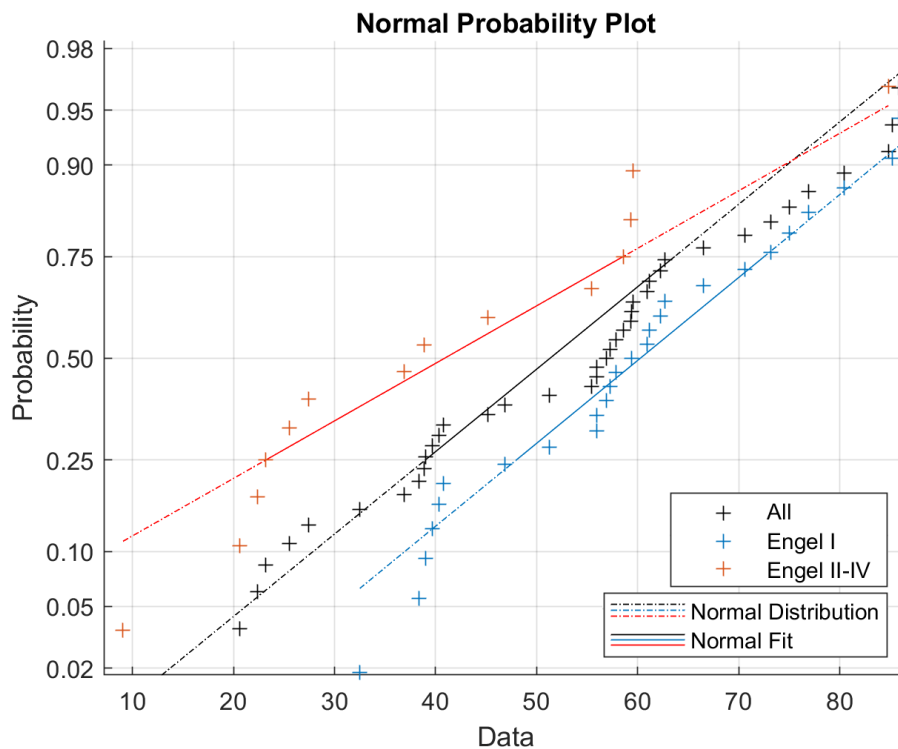
#### **3.4.4 Statistical Hypothesis**

Based on the resection indexes (RI) and the patient surgery outcomes (Engel groups), it is possible, using the Wilcoxon rank-sum test, to calculate the statistical significance ( $p$ -value) and verify whether the RI shows correlation with the outcomes. The tested hypothesis is that the RI does have the same median for the surgery outcome groups. The outcomes were divided into two groups: Good outcome (Engel I) and poor outcome (Engel II-IV).

From the 52 patients, 6 of them still do not have outcome information, so the maximal amount of patients for this analysis would, theoretically, be 46. However, as mentioned before, not all patients have data for all the epochs, so the analysis for sleep and awake combined data is composed of 41 patients, for sleep there are 42 patients, and for awake there are 45 patients.

As a non-parametric test, the rank-sum test does not require a normal distribution of the samples but has an excellent performance either way. It also deals with small sample sizes, which is the case of this work. The distribution of the samples can be seen in Figure 24.

Figure 24 – Distribution of the samples compared to the normal distribution.



Source: From the author.

### 3.4.5 Definition of a Threshold

As described, the methodology of picking the clusters which show epileptical activities comprises choosing the maximal amount of valid ones. However, clusters with very low percentages have a minor impact on the resection index (RI). Determining a minimal threshold that maximizes the statistical significance ( $p$ -value) by sequentially ignoring the lowest activity clusters by percentage, will eventually lead to only significant clusters remaining.

### 3.4.6 Statistics and Graphs

After the calculation of the indexes and the  $p$ -value, additional statistics were computed to support and better describe the results. The statistics include mean, median, standard deviation, as well as effect size statistics that support the understanding of the  $p$ -value



(i.e., Cohen's  $U_3$ , Hedges'  $g$ , and  $\Omega^2$ ). Histograms and box plots were also generated to facilitate the visualization of results.

### 3.5 PREDICTIVE ANALYSIS

The development of a second data mining to detect patterns that may predict the Focal Cortical Dysplasia (FCD) type of patients is totally reliant on the results on the first analysis, if no correlation is detected, the results would suggest that the mined information is not meaningful to the conditions of the patients and no relevant information can be extracted.

However, the first part of the analysis showed that does exist a strong correlation among the mined data and patient's surgery outcome, so it is reasonable to verify whether or not it is possible to mine new hidden patterns using the previous data as a baseline.

The vigilance period with a stronger statistical correlation was chosen for this analysis, in this case, the combination of sleep and awake.

It is essential to notice that this further exploration of the results must be considered more of an experimental investigation, given that the classification algorithms require bigger sample sizes, usually more than 50, and more samples are not available by the time of this writing. Patients with FCD III were dropped from the analysis because the sample size for this group is too small, seven patients, reducing the group of study to just 28 people.

The first step of the process is selecting features that may be relevant for the detection of the FCD type. The nine selected features are described in Table 9:

Table 9 – Selected features for the classification models.

<b>Selected Features</b>	<b>Description</b>
<b>Cluster 1 Percentage (%)</b>	Percentage of detected clustered activity present in the first cluster, most significant.
<b>Cluster 1 Active Region (AR) channels</b>	Number of channels in cluster one with higher activity (active) computed using the k-means method.
<b>Cluster 1 Active Region (AR) channels (%)</b>	Number of channels in cluster one with higher activity (active) as a percentage of the total number of channels.
<b>Cluster Percentages Skewness</b>	Calculation of the skewness of the cluster distribution percentages.
<b>Cluster Percentages Kurtosis</b>	Calculation of the kurtosis of the cluster distribution percentages.
<b>Cluster 1 Maximal IED Rate</b>	Maximal detected IED rate in cluster 1.
<b>Skewness of Maximal IED Rate</b>	Calculation of skewness of the maximal IED rate of all clusters.
<b>Kurtosis of Maximal IED Rate</b>	Calculation of kurtosis of the maximal IED rate of all clusters.
<b>Seizure Onset Zone Index – SOZI (%)</b>	Calculation of the percentage of the clustered activity inside the clinically defined SOZ region.

Source: From the author.

In possession of all information, it is possible to calculate the Seizure Onset Zone Index (SOZI). The index provides information about the overall IEDs detected by the channels, weighted by cluster activity, included in the clinically defined SOZ area. This new index is based on the calculus of the resection index and is defined by the formula below:

$$\mathbf{SOZI} = \sum_{cl} W_{cl} A_{cl}$$

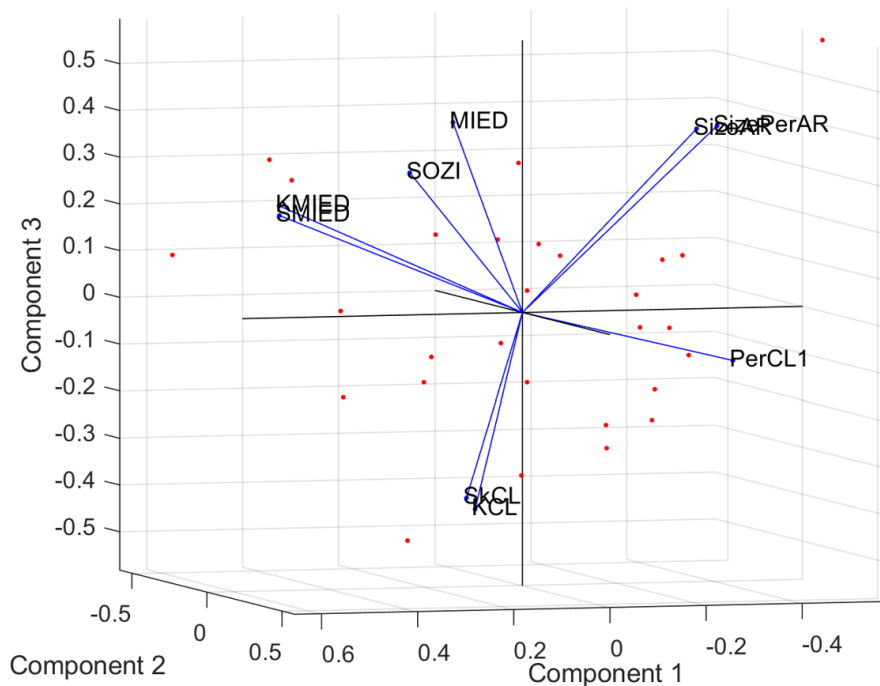
Where  $A_{cl}$  is the weight of a cluster, the amount of detected activities in a cluster as a percentage of the total detected activity. And  $W_{cl}$  is defined as:

$$W_{cl} = \frac{\sum_{ch \in SOZ} Q_{ch}}{\sum_{ch} Q_{ch}}$$

$W_{cl}$  represents the sum of the detected activities ( $Q$ ) of each channel inside the seizure onset zone area (SOZ), divided by the sum of detected activities of all channels, for a cluster.

The next step was to calculate the principal components using PCA. The results showed that five PCs can explain 95% of data variance. A biplot graph was used to plot the 3 PCA coefficients with higher significance (38.0%, 28.3%, and 19%), along with the variables as shown in Figure 25. The graphic representation revealed that some variables could be considered redundant, while others have a low contribution to the principal components.

Figure 25 – 3D plot showing the contribution of each feature in principal components.



*Legend:*

*MIED – Maximal IED Rate / KMIED – Kurtosis of MIED / SMIED – Skewness of MIED / SOZI – Seizure Onset Zone Index / SkCL – Skewness of Cluster Percentages / PerCL1 – Percentage of Cluster 1 / SizeAR – Size of Active Region / SizePerAR – Size of Active Region (%)*

Source: From the author.

Table 10 lists the five most relevant features that are used as the inputs of the classification models.

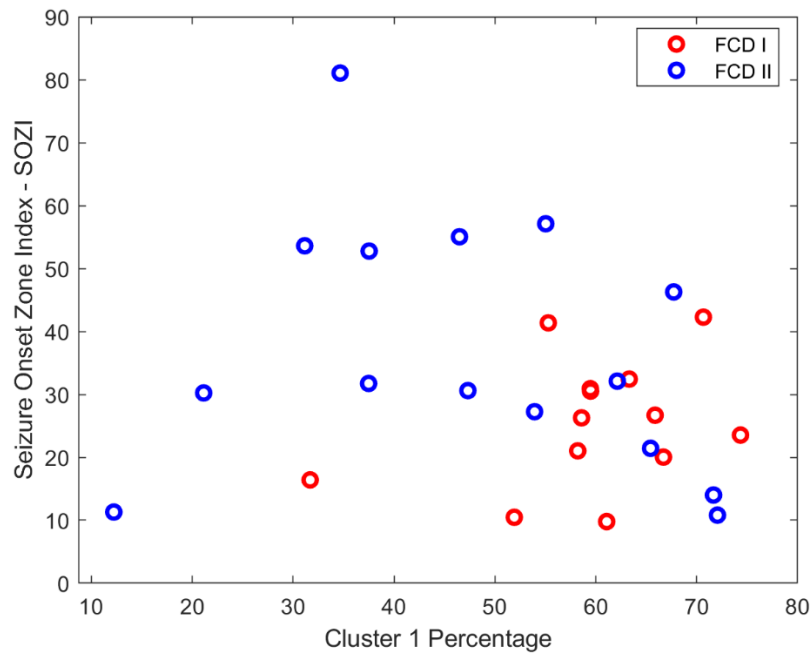
Table 10 – Selected features for the predictive analysis.

Most Relevant Features
Cluster 1 Percentage (%)
Cluster 1 Active Region (AR) channels
Cluster Percentage Skewness
Cluster 1 Maximal IED Rate
Seizure Onset Zone Index – SOZI (%)

Source: From the author.

Moreover, from these five features, plotting the samples of each group showed the two most significant for explaining FCD type groups are the “Cluster 1 percentage”, and the “seizure onset zone index” (SOZI). This can be easily noticed in a scatterplot (Figure 26).

Figure 26 – Scatterplot of FCD I and II for the two most significant features.



Source: From the author.

From this point on, it is possible to train the classification methods and refine the results, if possible.

Diverse methods were trained and compared, and the ones with consistently higher accuracy were the Ensemble and *k*NN, respectively. The Ensemble models also use *k*NN for its

weak learners. Additionally, tests with a reduced set of predictors were also investigated but provided worse results. Increasing the number of learning cycles did not improve the results over the accuracy achieved with 30 cycles.

Due to the low number of samples, the 5-fold cross-validation method was chosen instead of the 10-fold. The cross-validation is a process that randomly partitions the data into  $k$  sets, and for each set, reserve it as the test group and train the model using the other  $k-1$  groups. The accuracy is measured by the average of the saved training results. This method of validation is the most appropriate for a low number of samples.

As a form to verify the model's accuracy, the best approach is to test the model on new data that were not used in the test groups. The availability of six patients that were excluded from the statistical analysis due to lack of outcome information opened the chance to test the models on untested data that was chosen unintentionally. A late new batch of information of FCD type for five patients, that were still missing the report, increased the test group to eleven patients. This turned possible to refine the models adding the patients with wrong predictions and train the models once more with the prospects to increase the accuracy of the models.

Due to the fact that FCD III may be also classified as a dual pathology (FCD I/II + associated pathology), it opens the opportunity to test how the model performs in this state. The tests were done, and the results were also reported in the next chapter.

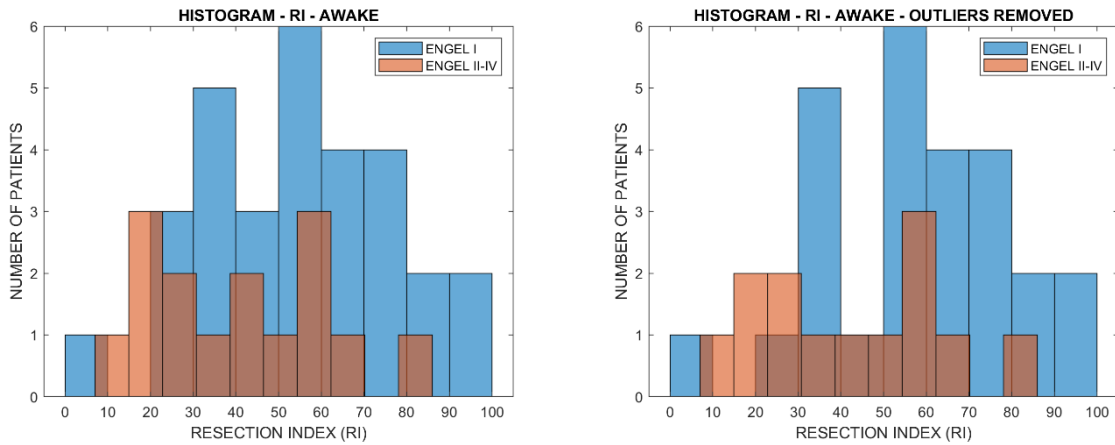
## **4 RESULTS AND DISCUSSION**

The following sub-sections will report the results, followed by a brief description of the presented data. In the end, the results will be discussed and confronted with other parts of the analysis or similar works.

### **4.1 AWAKE**

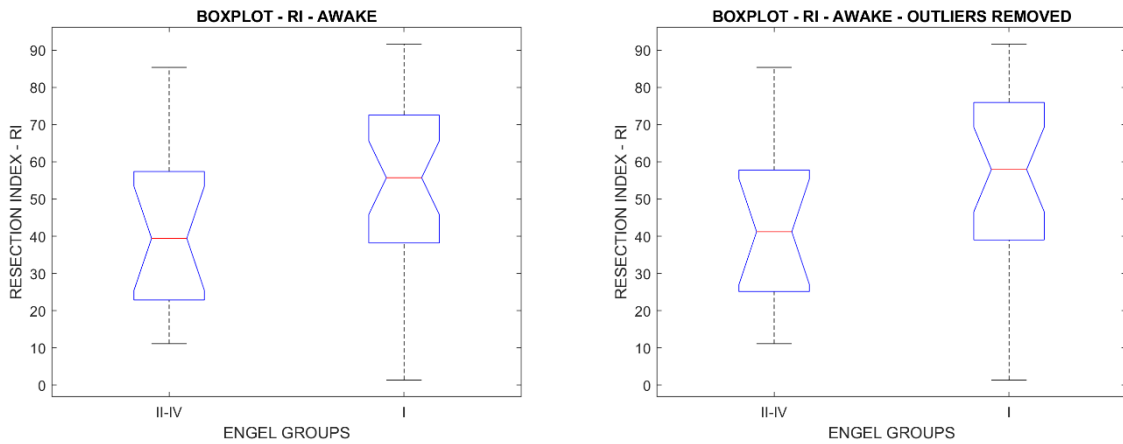
Figures 27, 28 and Table 11 display the graphs and the results of the awake data analysis in two scenarios, with (left) and without outliers (right).

Figure 27 – Histogram of the outcome groups for awake data.



Source: From the author.

Figure 28 – Box plot of the outcome groups for awake data.



Source: From the author.

Both histogram and boxplot show differences between Engel groups (outcomes), but they are moderate. The notches partly overlap, showing that the probability of the groups having different medians is less than 95%.

The group with Engel II-IV (poor outcome) produces a lower resection index median than the Engel I group, with good outcomes.

Graphically, the removal of outliers did not show much effect on the data.

Table 11 – Resection index statistics - Awake

STATISTICS	ALL DATA	WITHOUT OUTLIERS
<b>Number of patients</b>	<b>45</b>	<b>39</b>
Mean	50.3179	53.2554
Median	52.7198	56.1036
Std. Deviation	22.3585	22.3015
<b>Number of Engel I</b>	<b>30</b>	<b>26</b>
Mean	55.1622	58.5458
Median	55.7221	57.9369
Std. Deviation	21.6543	21.1213
<b>Number of Engel II-IV</b>	<b>15</b>	<b>13</b>
Mean	40.6293	42.6747
Median	39.4465	41.2048
Std. Deviation	20.5205	20.7828
<b><i>p</i>-Value</b>	<b>.04438</b>	<b>.03840</b>
Mean difference	14.5329	15.8711
Cohen's U3 (Conf. 95%)	0.6667	0.7692
Hedges' g (Conf. 95%)	0.6558	0.7208
Omega <sup>2</sup> (Conf. 95%)	0.0713	0.0865

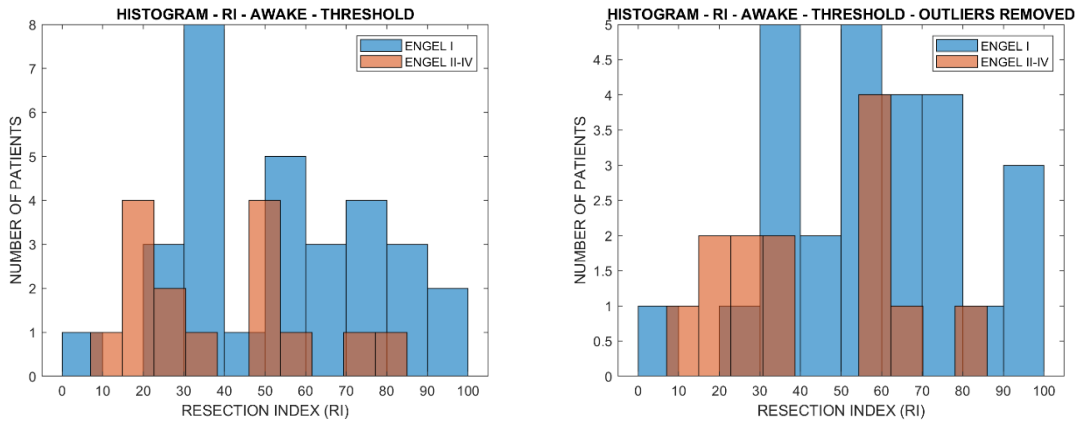
Source: From the author.

The *p*-value results indicate statistical significance ( $p < .05$ ), rejecting the hypothesis that the resection index (RI) does not correlate with the Engel groups.

These results suggest that the RI correlates with the Engel groups for good and poor outcomes when using the awake data for the analysis. The removal of outliers improved the statistical significance, also improving the scores of the effect size statistics.

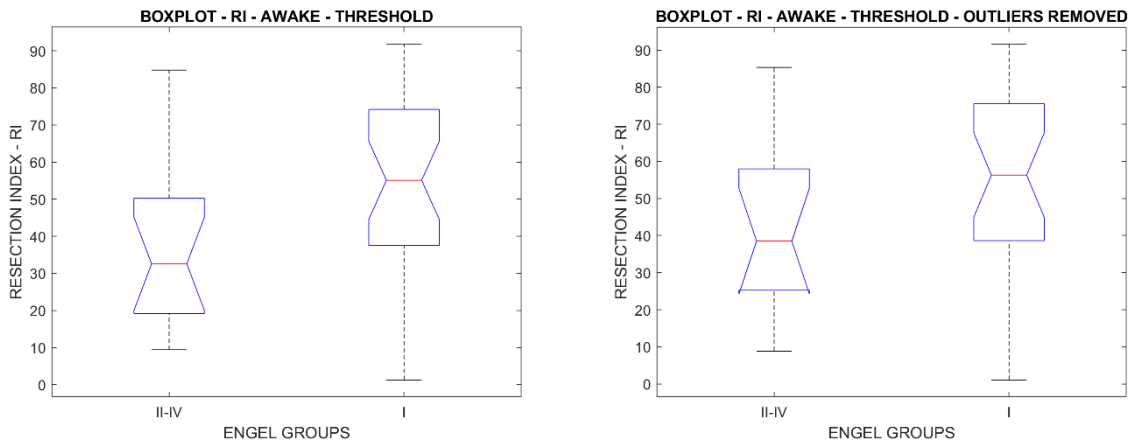
Figures 29, 30 and Table 12 display the graphs and the results of the awake data analysis, considering the calculated threshold, in two scenarios, with (left) and without the removal of outliers (right).

Figure 29 – Histogram of the outcome groups for awake data with threshold.



Source: From the author.

Figure 30 – Box plot of the outcome groups for awake data with threshold.



Source: From the author.

Both histogram and boxplot show differences between Engel groups and they are more prominent than the previous one. The notches partly overlap, showing that the probability of the groups having different medians is less than 95%.

The group with Engel II-IV (poor outcome) produces a lower RI median than the Engel I group, with good outcomes.

Visually, the removal of outliers did not show much effect on the data.



Table 12 – Resection index statistics with cluster threshold - Awake

STATISTICS	ALL DATA	WITHOUT OUTLIERS
Threshold	16%	5%
<b>Number of patients</b>	<b>45</b>	<b>39</b>
Mean	50.2564	52.0694
Median	54.5344	50.2829
Std. Deviation	23.0346	23.3625
<b>Number of Engel I</b>	<b>30</b>	<b>26</b>
Mean	55.1659	56.7589
Median	55.9042	56.0350
Std. Deviation	22.4159	23.2704
<b>Number of Engel II-IV</b>	<b>15</b>	<b>13</b>
Mean	40.4374	42.6905
Median	38.3307	47.3699
Std. Deviation	21.0288	20.5536
<b><i>p</i>-Value</b>	<b>.03117</b>	<b>.03570</b>
Mean difference	15.4048	16.0251
Cohen's U3 (Conf. 95%)	0.8000	0.6923
Hedges' g (Conf. 95%)	0.6454	0.7101
Omega <sup>2</sup> (Conf. 95%)	0.0686	0.0835

Source: From the author.

The *p*-value results show statistical significance ( $p < .05$ ), confirming the differences between groups.

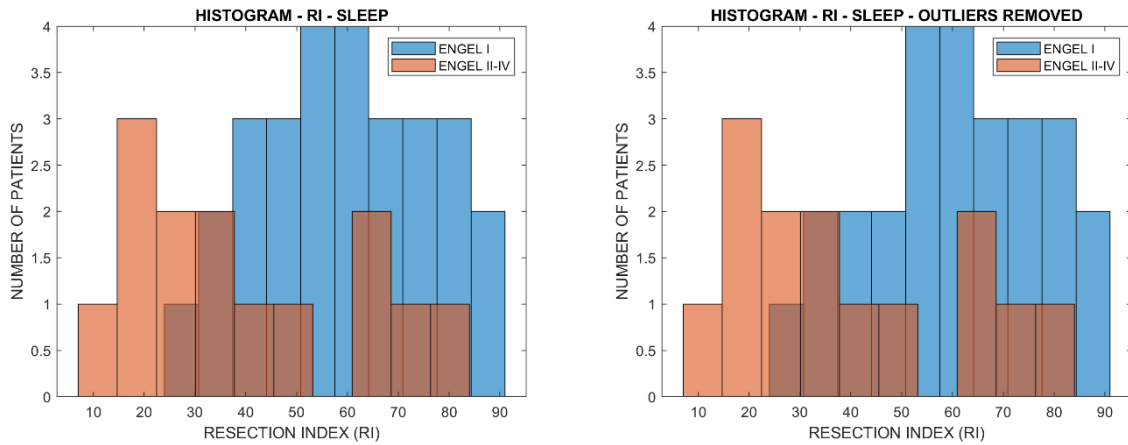
These results suggest that the resection index correlates with the Engel groups, for poor and good outcomes, using the awake data with a cluster threshold.

The removal of outliers improved the statistical significance, also improving the scores of the effect size statistics, except for Cohen's U3. The removal of the smallest clusters below a calculated threshold (16% and 5%, respectively) also displayed the effect of improving the statistical significance of the results.

#### 4.2 SLEEP

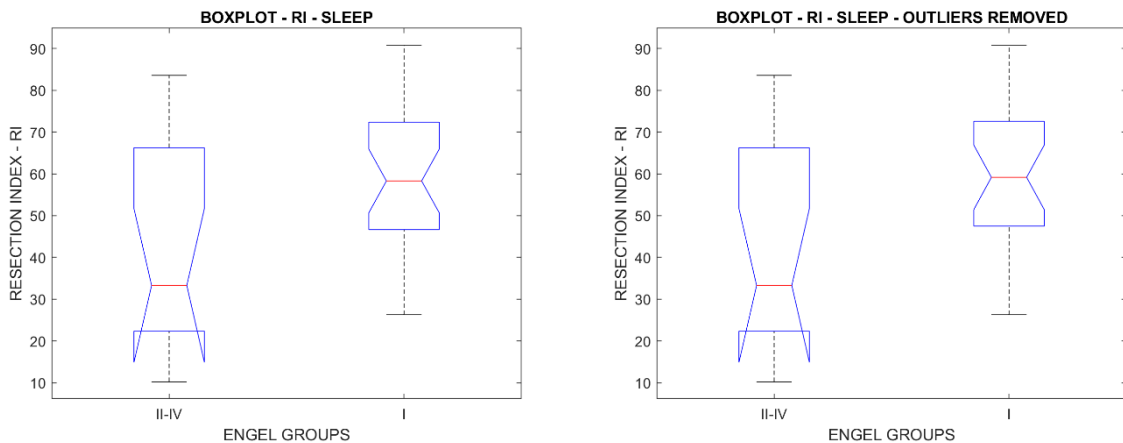
Figures 31, 32 and Table 13 display the graphs and data of the sleep data analysis in two scenarios, with (left) and without outliers (right).

Figure 31 – Histogram of the outcome groups for sleep data.



Source: From the author.

Figure 32 – Box plot of the outcome groups for sleep data.



Source: From the author.

Both histogram and boxplot show noticeable differences between groups that endorse the statistical analysis results. The notches do not overlap, indicating that there is 95% confidence that the samples are from groups with different medians, which is easily noticeable. The folded bottom of the Engel II-IV groups indicates that the notch is larger than the interquartile range (IQR).

The removal of outliers did not provide noticeable visual variation in the box plots.

Table 13 – Resection index statistics – Sleep

<b>STATISTICS</b>	<b>ALL DATA</b>	<b>OUTLIERS REMOVED</b>
<b>Number of patients</b>	<b>42</b>	<b>40</b>
Mean	53.3463	53.6963
Median	55.8129	55.9187
Std. Deviation	20.6013	21.0273
<b>Number of Engel I</b>	<b>28</b>	<b>26</b>
Mean	59.2686	60.2627
Median	58.2983	58.9275
Std. Deviation	16.4963	16.6680
<b>Number of Engel II-IV</b>	<b>14</b>	<b>14</b>
Mean	41.5017	41.5017
Median	33.2975	33.2975
Std. Deviation	23.6185	22.0995
<b><i>p</i>-Value</b>	<b>.01575</b>	<b>.01310</b>
Mean difference	17.7669	18.7610
Cohen's U3 (Conf. 95%)	0.7143	0.7143
Hedges' g (Conf. 95%)	0.9038	0.9419
Omega <sup>2</sup> (Conf. 95%)	0.1415	0.1562

Source: From the author.

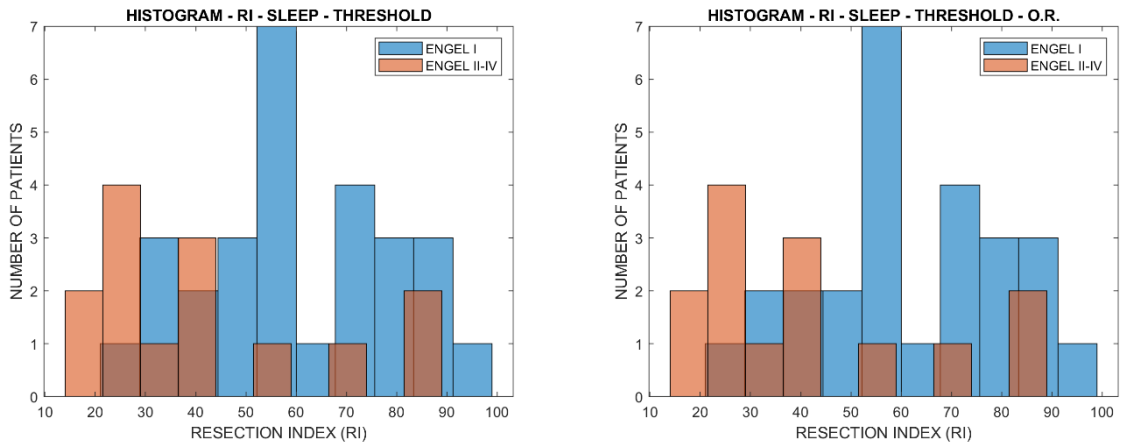
The *p*-value results show strong statistical significance ( $p < .05$ ), confirming the differences between groups.

These results suggest that the resection index correlates with the Engel groups, for poor and good outcomes, using the sleep data. In comparison to the awake vigilance state, the sleep state displays stronger statistical significance. The removal of outliers increased the statistical significance marginally, while slightly improving the scores of the effect size statistics.

The Hedges' *g* value indicates a large effect, with 0.9 and 0.94 S.D. between groups in the two scenarios.

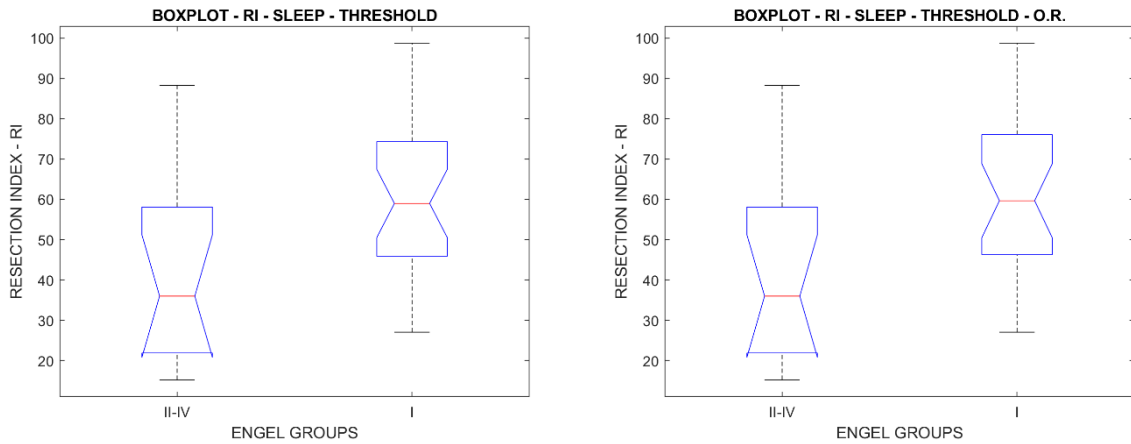
Figures 33, 34 and Table 14 display the graphs and the results of the sleep data analysis, considering the calculated threshold, in two scenarios, with (left) and without outliers (right).

Figure 33 – Histogram of the outcome groups for sleep data with threshold.



Source: From the author.

Figure 34 – Box plot of the outcome groups for sleep data with threshold.



Source: From the author.

Both histogram and boxplot show noticeable differences between groups that endorse the statistical analysis results. The notches do not overlap, indicating that there is 95% confidence that the samples are from groups with different medians.

Table 14 – Resection index statistics with cluster threshold – Sleep

STATISTICS	ALL DATA	OUTLIERS REMOVED
Threshold	14%	14%
<b>Number of patients</b>	<b>42</b>	<b>40</b>
Mean	54.2112	54.8079
Median	53.0255	54.6250
Std. Deviation	22.0656	22.3927
<b>Number of Engel I</b>	<b>28</b>	<b>26</b>
Mean	60.4241	61.8200
Median	58.9842	59.5277
Std. Deviation	18.7312	18.6277
<b>Number of Engel II-IV</b>	<b>14</b>	<b>14</b>
Mean	41.7856	41.7856
Median	35.9971	35.9971
Std. Deviation	21.3658	19.9339
<b><i>p</i>-Value</b>	<b>.00793</b>	<b>.00621</b>
Mean difference	18.6385	20.0344
Cohen's U3 (Conf. 95%)	0.7857	0.7857
Hedges' g (Conf. 95%)	0.8817	0.9451
Omega <sup>2</sup> (Conf. 95%)	0.1347	0.1572

Source: From the author.

The *p*-value results show statistical significance ( $p < .05$ ), attesting the existence of differences between groups.

These results suggest that the resection index correlates with the Engel groups, for poor and good outcomes, using the sleep data with a 14% threshold. The removal of outliers improved the statistical significance, while also improving the scores of the effect size statistics.

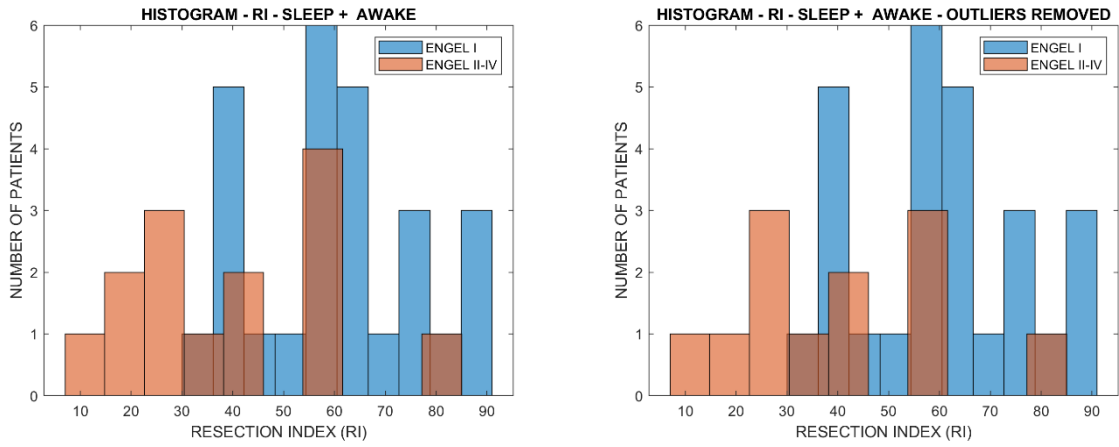
The removal of the smallest clusters below a calculated threshold (14% for both cases) drastically improved the statistical significance of the results.

Hedges' *g* values reveal the existence of a large effect, with the groups exhibiting 0.88 and 0.94 S.D. between them, in each scenario.

#### 4.3 SLEEP AND AWAKE

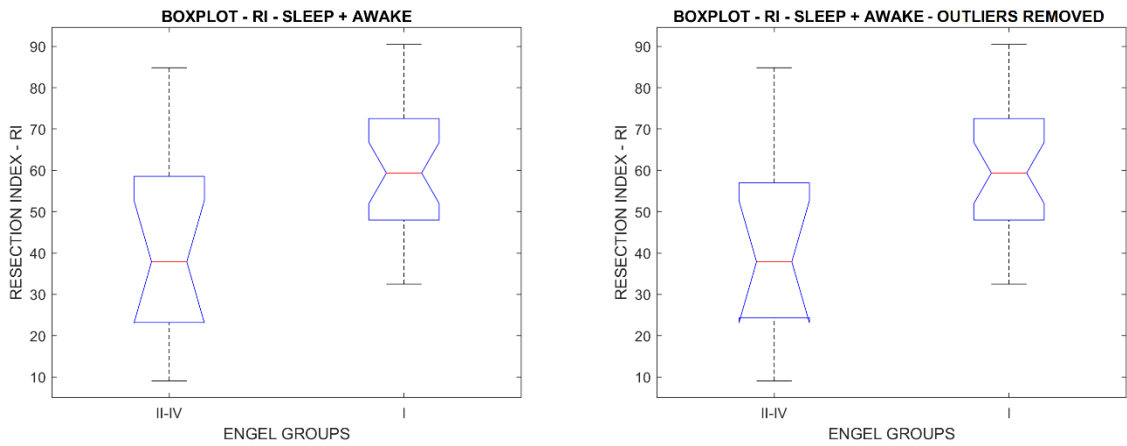
Figures 35, 36 and Table 15 display the graphs and data of the combination of sleep and awake data analysis in two scenarios, with (left) and without (right) of outliers. The combination of the epochs is also referred in the results as "Sleep + Awake".

Figure 35 – Histogram of the outcome groups for sleep and awake data.



Source: From the author.

Figure 36 – Box plot of the outcome groups for sleep and awake data.



Source: From the author.

Both histogram and boxplot show noticeable differences between groups that support the statistical analysis results. The notches do not overlap, indicating that there is 95% confidence that the samples are from groups with different medians.

Table 15 – Resection index statistics - Sleep + Awake

STATISTICS	ALL DATA	OUTLIERS REMOVED
<b>Number of patients</b>	<b>41</b>	<b>39</b>
Mean	53.4214	54.1058
Median	56.9295	56.9295
Std. Deviation	19.6799	19.4463
<b>Number of Engel I</b>	<b>27</b>	<b>27</b>
Mean	60.1275	60.1275
Median	59.3657	59.3657
Std. Deviation	15.6980	15.6980
<b>Number of Engel II-IV</b>	<b>14</b>	<b>12</b>
Mean	40.4884	40.5572
Median	37.8831	37.8831
Std. Deviation	20.1245	20.2336
<b><i>p</i>-Value</b>	<b>.00312</b>	<b>.00403</b>
Mean difference	19.6390	19.5703
Cohen's U3 (Conf. 95%)	0.8571	0.9167
Hedges' g (Conf. 95%)	1.0834	1.0843
Omega <sup>2</sup> (Conf. 95%)	0.2001	0.1905

Source: From the author.

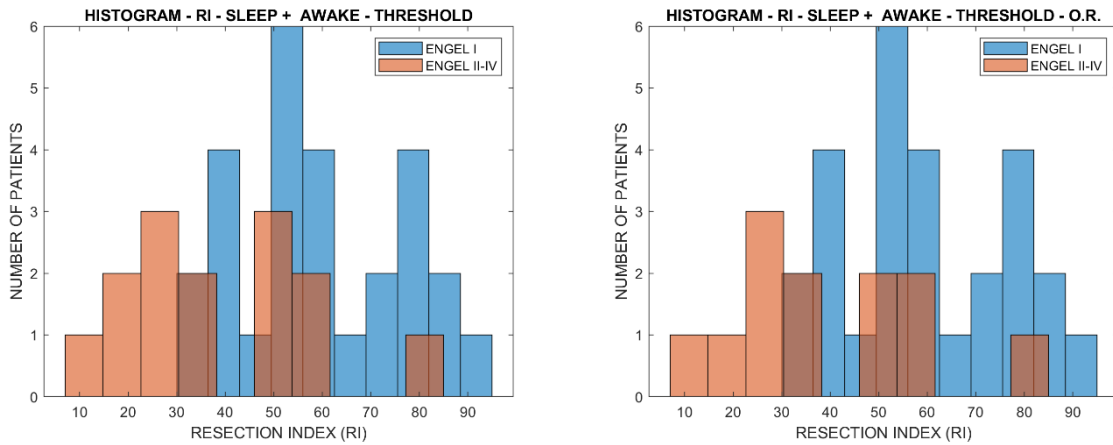
The *p*-value results represent strong statistical significance ( $p < .05$ ), indicating the existence of differences between groups.

These results suggest that the resection index shows a strong correlation with the Engel groups, for poor and good outcomes, for the combination of sleep and awake data. The removal of outliers decreased the statistical significance but improved the results of some effect size statistics, especially for the Cohen's U3, meaning that 91.67% of the samples of Engel II-IV (poor outcome) are below the median of the group with Engel I (good outcome).

Omega<sup>2</sup> values represent that 20% and 19% of the variance in the resection index is explained by the Engel group membership.

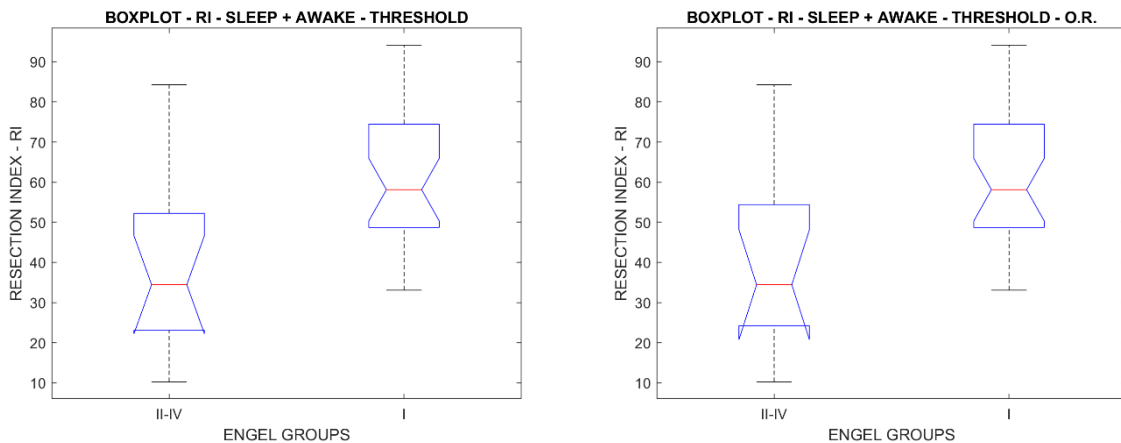
Figures 37, 38 and Table 16 display the graphs and data of the sleep and awake with the removal of clusters below a determined threshold. Two scenarios are analyzed, with (left) and without outliers (right).

Figure 37 – Histogram of the outcome groups for sleep and awake data with threshold.



Source: From the author.

Figure 38 – Box plot of the outcome groups for sleep and awake data with threshold.



Source: From the author.

Both histogram and boxplot visually show differences between groups that support the statistical analysis results. The notches do not overlap, indicating that there is 95% confidence that the samples are from groups with different medians.



Table 16 – Resection index statistics with cluster threshold - Sleep + Awake

STATISTICS	ALL DATA	OUTLIERS REMOVED
Threshold	10%	10%
<b>Number of patients</b>	<b>41</b>	<b>39</b>
Mean	52.9419	53.8855
Median	53.2154	53.6025
Std. Deviation	20.2215	20.0423
<b>Number of Engel I</b>	<b>27</b>	<b>27</b>
Mean	60.1706	60.1706
Median	58.1186	58.1186
Std. Deviation	16.5409	16.5409
<b>Number of Engel II-IV</b>	<b>14</b>	<b>12</b>
Mean	39.0010	39.7440
Median	34.4482	34.4483
Std. Deviation	19.3580	20.0260
<b><i>p</i>-Value</b>	<b>.00198</b>	<b>.00488</b>
Mean difference	21.1696	20.42650
Cohen's U3 (Conf. 95%)	0.9286	0.9167
Hedges' g (Conf. 95%)	1.1534	1.10200
Omega <sup>2</sup> (Conf. 95%)	0.2228	0.19610

Source: From the author.

The *p*-value results represent strong statistical significance ( $p < .05$ ), rejecting the hypothesis that the samples are from groups with the same median.

This means that the resection index, when clusters with activity lower than 10% are removed, provided the best correlation with the Engel groups (poor and good outcome groups) with a *p*-value of .00198. This result also showed a stronger correlation compared to any of the previous analysis. Indicating that the combined state of sleep and awake vigilance epochs would be the most relevant state to extract information.

Like the previous result, the removal of outliers decreased the statistical significance, but this time also decreased the significance of the effect size statistics.

All the effect size statistics showed strong effects for these scenarios. Cohen's U3 indicates that 92.8% and 91.6% of patients in the group with poor outcomes are below the median of the group with good outcomes. Hedges' g reveals that the S.D. between groups is 1.15 and 1.10 for each scenario, respectively. And the Omega<sup>2</sup> results represent that 22.2% and 19.6% of the variance of the RI can be explained by the Engel group membership.

#### 4.4 DISCUSSION OF THE STATISTICAL RESULTS

The statistical analysis indicates that the epochs show different significance weights with the combined state of sleep and awake, having significantly better correlation than the other two states. Additionally, between awake and sleep, the sleep epoch reveals to have more relevance than awake; this result is consistent with a brand new related study findings (PETR KLIMES et al., 2019). According to this same study, though, the NREM sleep state performs better than the combined states, which disagrees with the results of this work. However, the studies are not entirely comparable since, in this related research, the sleep epoch was split and analyzed by its cycles.

The fact that the separate awake and sleep vigilance states demonstrate lower significance than when combined, suggests that these vigilance periods are complementary, and the brain regions generating IEDs in one might be different in the other, which seems to corroborate with the theory. The combined state proved to be the most relevant to the extraction of information and might be the most appropriate vigilance state to be evaluated in order to better delineate the epileptogenic zone prior to the resective surgery.

The results also demonstrated that the employed data mining algorithm excels in extracting relevant information and in grouping the detected activities, however, the fact that a definition of a cluster threshold improved the results for all the analysis, indicates that the clustering process can be improved. Particularly because the alternative coefficient of separability setting (low separability) did not perform well in a previous study (INÁCIO; JANCA, 2019). The solution might be finding a consistent approach between these two coefficients.

Finally, due to the significance of the results, the combined awake and sleep vigilance periods were taken as the basis for training the classification models. The results of this analysis are described in the next subsection.

#### 4.5 CLASSIFICATION RESULTS

The next subsections will describe and present the results of the classification models trained to predict the FCD groups of patients, using five features as input.

#### 4.5.1 Ensemble (Subspace $k$ NN)

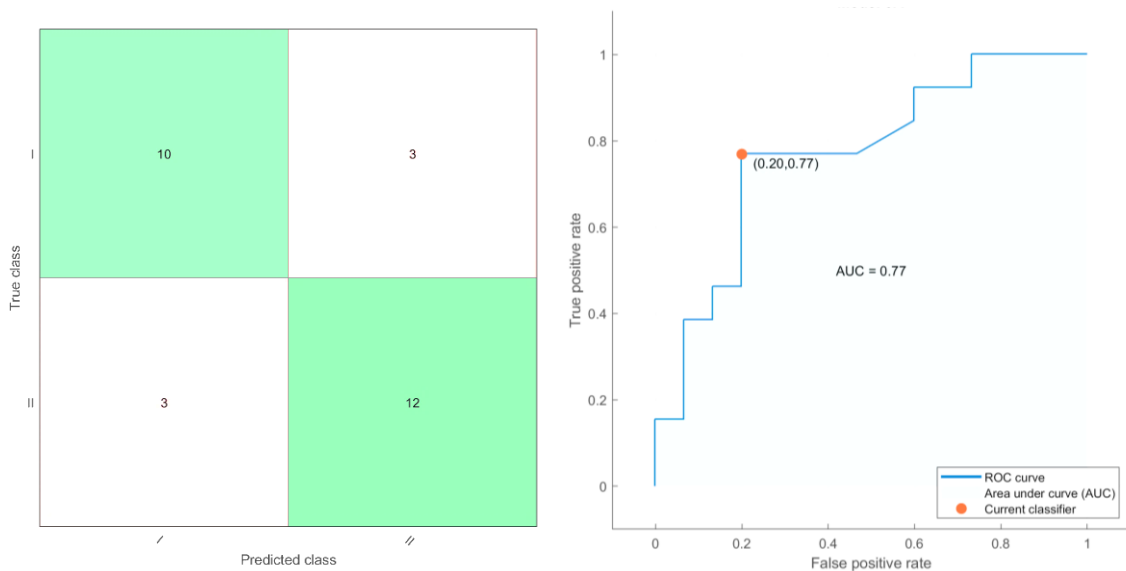
Below are described the parameters of training and the performance indicators.

Model parameters:

- **Method:** Subspace
- **Learning Cycles:** 30
- **Learners:**  $k$ NN
- **Combination of Predictors:** 3 subspace dimensions
- **Average accuracy:** 78.6%
- **Prediction speed:** ~40 obs/sec
- **Training time:** 4.2s

Figures 39 and 40 show the confusion matrices and the ROC curve of the trained model.

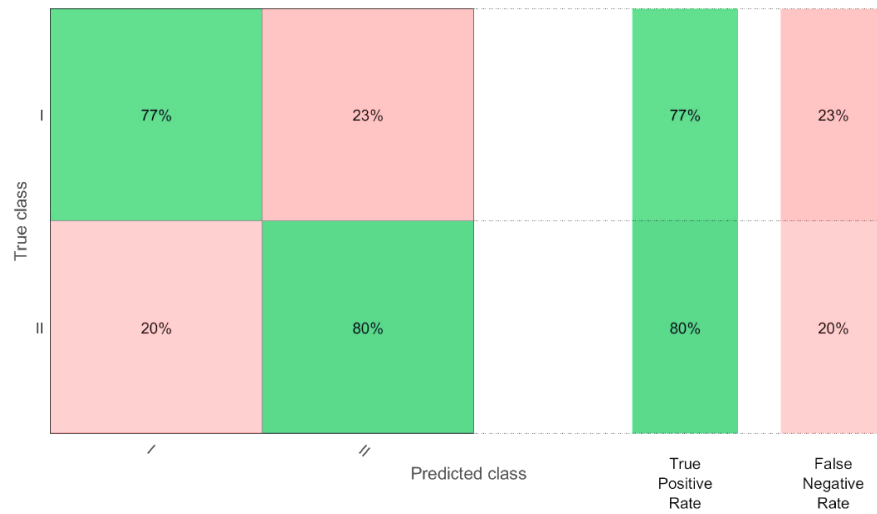
Figure 39 – Confusion matrix and ROC plot for the Ensemble model.



Source: From the author.

The ROC curve shows the performance of the model at all thresholds. The AUC of 0.77 indicates that the chance of the rank of a positive example being greater than a negative one is 77%, which is a measure of the quality of the model's predictions.

Figure 40 – Confusion matrix with true positive and false negative rates.



Source: From the author.

Even though the training model offers an average accuracy, the best method for verifying the model's accuracy is testing its results with a test group composed of data that was not used for training.

The group of 11 patients below was chosen by chance. Six of them were not employed for the first statistical analysis for lacking outcome data, however, they can be included in the test group for this analysis since they have FCD type info. For the other five, the information on their FCD types was obtained after the generation of the model, so they were included in the test group. The prediction results for this group of patients are presented in Table 17.

Table 17 – Ensemble prediction table for new patients.

PATIENT	FCD TYPE	PREDICTED GROUP
P143	I	I
P150	II	II
P163	II	II
P177	I	I
P179	I	I
P185	II	II
P025	II	I
P034	I	II
P043	II	II
P091	II	II
P117	II	II
<b>Correct Predictions</b>		<b>9 of 11 (81.82%)</b>

Source: From the author.

An additional test with the FCD III patients was done in order to test the model's accuracy further. The predicted group was compared to the dual pathology reclassification of FCD III, the results are presented in Table 18.

Table 18 – Ensemble prediction for the FCD III patients, reclassified as dual pathology.

PATIENT	FCD TYPE	FCD DUAL PATHOLOGY	PREDICTED GROUP
P048	III	I	I
P085	III	II	II
P096	III	I	I
P129	III	I	I
P136	III	I	I
P142	III	I	I
P173	III	I	II
<b>Correct Predictions</b>			<b>6 of 7 (85.71%)</b>

Source: From the author.

#### 4.5.2 Ensemble (Subspace $k$ NN) - Refined

Below are described the parameters of training and the performance indicators.

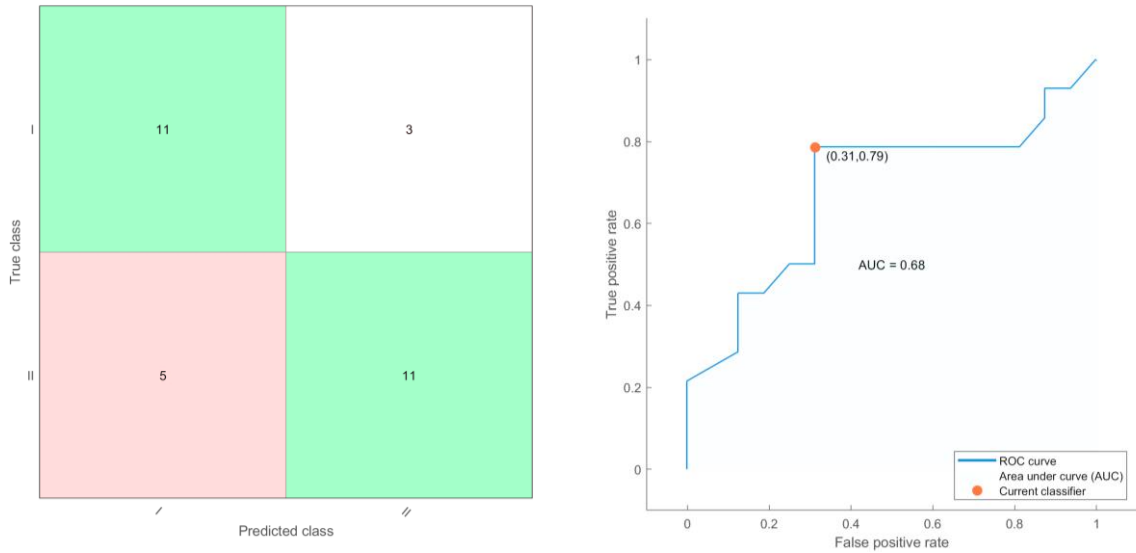
Model parameters:

- **Method:** Subspace
- **Learning Cycles:** 30
- **Learners:**  $k$ NN
- **Combination of Predictors:** 3 subspace dimensions
- **Average accuracy:** 73.3%
- **Prediction speed:** ~71 obs/sec
- **Training time:** 3.34s

One of the approaches to refine a classification model is to include new samples that had a false prediction for the training group. This is especially true in the case of this work, where the number of samples is considered below the ideal. Therefore, P025 and P034 were added to the training group and the model was retrained. The average accuracy dropped slightly when compared to the previous model, but the new model continued to display precise predictions to all the other nine patients (Table 19).

Figures 41 and 42 display the confusion matrices and the ROC plot for the refined model.

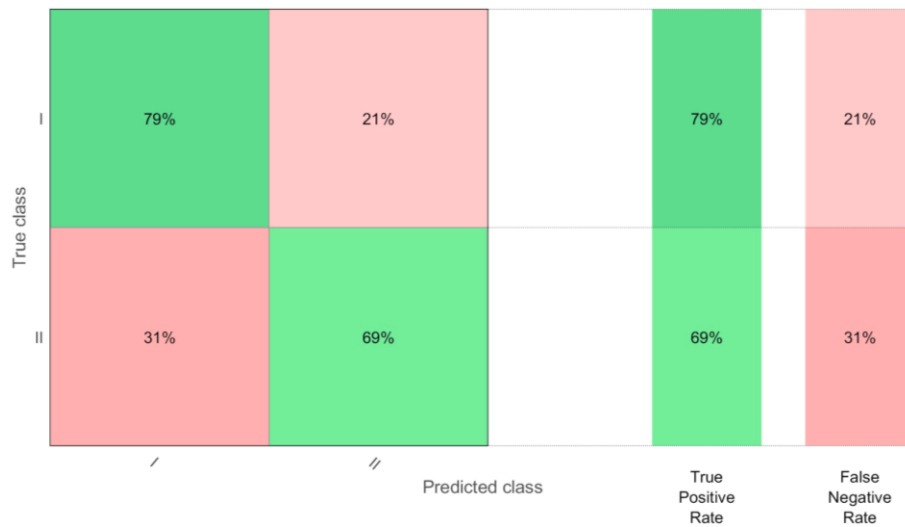
Figure 41 – Confusion matrix and ROC plot for the refined Ensemble model.



Source: From the author.

The AUC of 0.68 indicates that the chance of the rank of a positive example being greater than a negative one is 68%, measuring the overall quality of the model’s predictions.

Figure 42 – Confusion matrix with true positive and false negative rates.



Source: From the author.

Table 19 – Ensemble prediction table for new patients - Refined model.

PATIENT	FCD TYPE	PREDICTED GROUP
P143	I	I
P150	II	II
P163	II	II
P177	I	I
P179	I	I
P185	II	II
P043	II	II
P091	II	II
P117	II	II
<b>Correct Predictions</b>		<b>9 of 9 (100.0%)</b>

Source: From the author.

#### 4.6 DISCUSSION OF CLASSIFICATION RESULTS

The trained models demonstrated good and consistent results with all tests on untrained data showing accuracy superior to 80%, even for the FCD III type, where the medical classification is still imprecise.

Unfortunately, it is impossible to compare and discuss the results with a related study, because this type of approach of using machine learning methods on EEG data to predict the FCD types has never been done before. The obstacles and the difficult access to the EEG recordings in combination with other necessary medical data of patients are probably one of the reasons for the absence of this kind of analysis. As well as the fact that is an area that is still developing and being researched with less funding when compared to other diseases.

Despite the impossibility of comparison with third party results, the results reveal that it is possible to extract new patterns on the mined data from the iEEG analysis since these patterns extracted exhibited a strong relationship (from the predictions) to the actual medical data of the patients.

## 5 CONCLUSION

In this work two types of experiments were conducted, the first one presented a statistical analysis of iEEG recordings of patients with focal cortical dysplasia, using a data mining algorithm that detects, analyzes and groups IEDs by their spread patterns. The goal of the analysis was to investigate the association between the analyzed clustered data and the patient's surgery outcomes.

The conjunct of the algorithm's outputs and the medical information enables the calculation of a resection index and posterior inspection of statistical correlation. Additionally, examining the different vigilance periods, it was possible to identify the most relevant for analysis, which might offer more pertinent information for the better delineation of the epileptogenic zone, and therefore, optimizing the surgery resected area. These results partially come to an agreement, with a brand-new study, over the significance of the vigilance epochs.

Given the significance of the first results, which showed that the extracted data was meaningful, a second data mining was done in order to, using the algorithm's outputs, define parameters that help to predict the patient's FCD classes. The classification models trained in this experiment revealed satisfactory accuracy when tested on new data and in a distinct scenario. Given the unprecedented character of the analysis, comparisons with the accuracy of related works are unfeasible.

As future work, it would be reasonable to test if better accuracy is obtained with statistical or neural network methods. Besides, further investigation on how to identify the epileptogenic area is necessary, since the approach used in this work depends on the surgery results, and ideally, the identification of these areas should not be influenced by the surgery or clinician's decisions, which are also subject to errors.



## REFERENCES

ACM. A Short Introduction to K-Nearest Neighbors Algorithm. **Association for Computing Machinery Blog**, 2016. Available at: <<https://helloacm.com/a-short-introduction-to-k-nearest-neighbors-algorithm/>>. Accessed: 14 November 2019.

AGGARWAL, C. C. **Data Mining: The Textbook**. [S.l.]: Springer International Publishing, 2015. 734 p. ISBN 9783319141428.

ANEJA, S.; JAIN, P. Refractory Epilepsy in Children. **The Indian Journal of Pediatrics**, v. 81, n. 10, p. 1063–1072, August 2014.

AZEVEDO, A. Data Mining and Knowledge Discovery in Databases. In: MEHDI KHOSROW-POUR, D. B. A. **Encyclopedia of Information Science and Technology**. 4th. ed. [S.l.]: IGI Global, 2017. Cap. 166, p. 1907-1918.

AZEVEDO, A.; SANTOS, M. F. **KDD, SEMMA and CRISP-DM: A parallel overview**. IADIS European Conference on Data Mining. Amsterdam: DBLP. 2008. p. 24-26.

BAE, Y.-S. et al. New Classification of Focal Cortical Dysplasia: Application to Practical Diagnosis. **Journal of Epilepsy Research**, v. 2, n. 2, p. 38-42, December 2012.

BAGHERI, E. et al. Interictal epileptiform discharge characteristics underlying expert interrater agreement. **Clinical Neurophysiology**, v. 128, n. 10, p. 1994-2005, October 2017.

BANERJEE, P. N.; HAUSER, W. A. Incidence and Prevalence. In: JEROME ENGEL, T. A. P. J. A. **Epilepsy: A Comprehensive Textbook**. [S.l.]: Lippincott Williams & Wilkins, v. 3, 2008. Cap. 5, p. 45-56. ISBN 9780781757775.

BARTOLOMEI, F. et al. Defining epileptogenic networks: Contribution of SEEG and signal analysis. **Epilepsia**, v. 58, n. 7, p. 1131-1147, July 2017.

BERRY, M. J. A.; LINOFF, G. S. **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management**. 2. ed. [S.l.]: Wiley, 2004.

BESBEAS, P. ESTIMATION. In: SALKIND, N. J. **Encyclopedia of Research Design**. [S.l.]: SAGE Publications Inc., 2012. p. 419-422.

BLÜMCKE, I. et al. The clinico-pathological spectrum of Focal Cortical Dysplasias: a consensus classification proposed by an ad hoc Task Force of the ILAE Diagnostic Methods Commission. **Epilepsia**, v. 52, n. 1, p. 158-174, January 2011.

BRUNO, E. et al. Epilepsy and Neurocysticercosis in Latin America: A Systematic Review and Meta-analysis. **PLoS Neglected Tropical Diseases**, v. 7, n. 10, 31 October 2013.

CAPRARO, R. M.; YETKINER, Z. E. p VALUE. In: SALKIND, N. J. **Encyclopedia of Research Design**. [S.l.]: SAGE Publications, Inc., 2012. p. 1180-1184.

CHAPMAN, P. et al. **CRISP-DM 1.0 Step-by-step data mining guides**. SPSS. [S.l.], p. 76. 2000.

CURTIS, M. D. et al. Interictal Epileptiform Discharges in Partial Epilepsy: Complex Neurobiological Mechanisms Based on Experimental and Clinical Evidence. In: NOEBELS JL, A. M. R. M. E. A. **Jasper's Basic Mechanisms of the Epilepsies**. 4th. ed. [S.l.]: Bethesda (MD), National Center for Biotechnology Information (US), 2012.

DEWOLFE, J. L.; MALOW, B. A. Therapy in Sleep Medicine. In: BARKOUKIS, T. J. et al. **Therapy in Sleep Medicine**. [S.l.]: Saunders, 2012. Cap. 50, p. 629-646.

DOYLE, D. Notched Box Plots. **David's Statistics**, 2016. Available at: <<https://sites.google.com/site/davidsstatistics/home/notched-box-plots>>. Accessed: 8 November 2019.

DRESCH, A. et al. **Design Science Research: Metodo de Pesquisa para Avanço da Ciência e Teconologia**. Porto Alegre: Bookman, 2015.

ENGEL, J. **Surgical Treatment of the Epilepsies**. 2nd. ed. [S.l.]: Lippincott Williams & Wilkins, 1993.

ENGEL, J. J.; PEDLEY, T. A. Introduction: What is epilepsy? In: JEROME ENGEL, T. A. P. J. A. **Epilepsy: A Comprehensive Textbook**. [S.l.]: Lippincott Williams & Wilkins, v. III, 2008. Cap. 1, p. 1 - 8. ISBN 9780781757775.

EPILEPSY ACTION AUSTRALIA. Nocturnal Seizures – Seizures during Sleep. **Epilepsy Action Australia**, 2017. Available at: <<https://www.epilepsy.org.au/about-epilepsy/understanding-epilepsy/nocturnal-seizures-seizures-during-sleep/>>. Accessed: 12 August 2019.

FAUSER, S. et al. Long-term seizure outcome in 211 patients with focal cortical dysplasia. **Epilepsia**, v. 56, n. 1, p. 66-76, January 2015.

FAY, D. S.; GEROW, K. A biologist's guide to statistical thinking and analysis. In: COMMUNIT, T. C. E. R. **WormBook**. [S.l.]: WormBook, 2013. p. 47-48.

FAYYAD, U. et al. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, 1996.

FEINDEL, W. et al. Epilepsy Surgery: Historical Highlights 1909–2009. **Epilepsia**, v. 50, n. s3, p. 131-151, March 2009.

FISHER, R. S. et al. Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). **Epilepsia**, v. 46, n. 4, p. 470-472, April 2005.

FISHER, R. S. et al. A practical clinical definition of epilepsy. **Epilepsia**, 55, n. 4, 2014. 475–482.

FISHER, R. S. et al. Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology. **Epilepsia**, v. 58, n. 4, p. 522-530, April 2017.

FLEXER, A. Data Mining and EEG, Vienna, June 2000.

FORD, C. The Wilcoxon Rank Sum Test. **University of Virginia Library**, 2017. Available at: <<https://data.library.virginia.edu/the-wilcoxon-rank-sum-test/>>. Accessed: 20 October 2019.

FRAWLEY, W. J. et al. Knowledge Discovery in Databases: An Overview. **AI Magazine**, v. 13, n. 3, p. 57-70, 1992.

FUNES, A.; DASSO, A. **Data Mining and the KDD Process**. 4th. ed. [S.l.]: Encyclopedia of Information Science and Technology, 2018.

GEORGIEV, G. Statistical Significance in A/B Testing – a Complete Guide. **Analytics-Toolkit**, 2018. Available at: <<http://blog.analytics-toolkit.com/2017/statistical-significance-ab-testing-complete-guide/>>. Accessed: 21 October 2019.

GLEN, S. p Value. **Statistics How To**, 2014. Available at: <<https://www.statisticshowto.datasciencecentral.com/p-value/>>. Accessed: 21 October 2019.

GLEN, S. Hedges' g. **Statistics How To**, 2016. Available at: <<https://www.statisticshowto.datasciencecentral.com/hedges-g/>>. Accessed: 21 October 2019.

GRABOWSKI, D. C. et al. Changing the neurology policy landscape in the United States: Misconceptions and facts about epilepsy. **Health Policy**, v. 122, n. 7, p. 797-802, July 2018.

GRINENKO, O. et al. A fingerprint of the epileptogenic zone in human epilepsies. **Brain**, v. 141, n. 1, p. 117-131, January 2018.

GUAN, J. et al. Surgical strategies for pediatric epilepsy. **Translational Pediatrics**, v. 5, n. 2, p. 55-66, April 2016.

HAN, J. et al. **Data Mining: Concepts and Techniques**. 3rd. ed. [S.l.]: Elsevier, 2011. 744 p.

HANEL, P.; MEHLER, D. Beyond Reporting Statistical Significance: Identifying Informative Effect Sizes to Improve Scientific Communication, 28 January 2018. 1-33.

HEDGES, L. V. Distribution theory for Glass's estimator of effect size and related estimators. **Journal of Educational Statistics**, v. 6, p. 107-128, 1981.

HENTSCHKE, H. Measures of Effect Size - Toolbox. **Github**, 2018. Available at: <[https://github.com/hhentschke/measures-of-effect-size-toolbox/blob/master/doc/Documentation\\_MESToolbox.pdf](https://github.com/hhentschke/measures-of-effect-size-toolbox/blob/master/doc/Documentation_MESToolbox.pdf)>. Accessed: 21 October 2019.

IBM. Background of KNN. **IBM Knowledge Center**, 2018. Available at: <[https://www.ibm.com/support/knowledgecenter/en/SS6NHC/com.ibm.swg.im.dashdb.analytics.doc/doc/r\\_knn\\_background.html](https://www.ibm.com/support/knowledgecenter/en/SS6NHC/com.ibm.swg.im.dashdb.analytics.doc/doc/r_knn_background.html)>. Accessed: 10 November 2019.

ILIADIS, L.; JAYNE, C. **Engineering Applications of Neural Networks**: 16th International Conference. International Conference on Engineering Applications of Neural Networks, EANN. Rhodes: Springer. 2015. p. 92-93.

INÁCIO, G. S.; JANCA, R. UNIGOU 2019: Final Posters. **INCBAC NEWS**, 2019. Available at: <<https://incbacnews.wordpress.com/2019/01/20/unigou-2019-final-posters/#jp-carousel-2880>>. Accessed: 14 June 2019.

JANCA, R. Analýzy intrakraniálního EEG. **SAMI - Signal Analysis, Modelling and Interpretation**, 2019. Available at:

<[http://sami.fel.cvut.cz/bsg/cv\\_iEEG/BSG\\_2018\\_web.pdf](http://sami.fel.cvut.cz/bsg/cv_iEEG/BSG_2018_web.pdf)>. Accessed: 10 August 2019.

JANCA, R. et al. Automatic detection and spatial clustering of interictal discharges in invasive recordings. **2013 IEEE International Symposium on Medical Measurements and Applications (MeMeA)**, 2013. ISSN 10.1109/MeMeA.2013.6549739.

JANCA, R. et al. Detection of Interictal Epileptiform Discharges Using Signal Envelope Distribution Modelling: Application to Epileptic and Non-Epileptic Intracranial Recordings. **Brain Topography**, v. 28, n. 1, p. 172-183, 2015. ISSN 15736792.

JANCA, R. et al. The sub-regional Functional Organization of neocortical irritative epileptic networks in Pediatric epilepsy. **Frontiers in Neurology**, v. 9, n. MAR, p. 1-11, March 2018. ISSN 16642295.

JEROME ENGEL, J. Approaches to refractory epilepsy. **Annals of Indian Academy of Neurology**, v. 17, n. 1, p. S12-S17, March 2014.

KABAT, J.; KRÓL, P. Focal cortical dysplasia - review. **Polish Journal of Radiology**, v. 77, n. 2, p. 35-43, 2012. ISSN 01377183.

KERR, M. P. The impact of epilepsy on patients' lives. **Acta Neurologica Scandinavica**, v. 126, n. s194, p. 1-9, October 2012.

KIM, S. B.; SUKCHOTRAT, T. DATA MINING. In: SALKIND, N. J. **Encyclopedia of Research Design**. [S.l.]: SAGE Publications, Inc., 2012. p. 353-356.

KIRIAKOPOULOS, E. Tests Before Surgery. **Epilepsy Foundation**, 2018. Available at: <<https://www.epilepsy.com/learn/treating-seizures-and-epilepsy/surgery/tests-surgery>>. Accessed: 20 August 2019.

KOVAC, S. et al. Invasive epilepsy surgery evaluation. **Seizure**, v. 44, p. 125-136, January 2017. ISSN 1059-1311.

KRAL, T. et al. Focal cortical dysplasia: long term seizure outcome after surgical treatment. **Journal of Neurology, Neurosurgery & Psychiatry**, v. 78, n. 8, p. 853–856, August 2007.

KWAN, P. et al. Definition of drug resistant epilepsy: Consensus proposal by the ad hoc Task Force of the ILAE Commission on Therapeutic Strategies. **Epilepsia**, v. 51, n. 6, p. 1069-1077, June 2010.

LACHAUX, J. P. et al. Intracranial EEG and human brain mapping. **Journal of Physiology Paris**, v. 97, p. 613-628, 2003. ISSN 09284257.

LAMORTE, W. W. Nonparametric Tests. **Boston University School of Public Health**, 2017. Available at: <[http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704\\_nonparametric/BS704\\_Nonparametric4.html](http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric4.html)>. Accessed: 30 October 2019.

LEVENTHAL, B. An introduction to data mining and other techniques for advanced analytics. **Journal of Direct, Data and Digital Marketing Practice**, v. 12, n. 2, p. 137-153, 2010. ISSN 17460166.

LEVER, J. et al. Principal component analysis. **Nature Methods**, v. 14, p. 641-642, June 2017.

LÜDERS, H. O. et al. The epileptogenic zone: general principles. **Epileptic Disord**, v. 8, n. Suppl. 2, p. S1-9, 2006.

MARKETS AND MARKETS. **Data Mining Tools Market by Component (Tools and Services), Business Function (Marketing, Finance, Supply Chain and Logistics, and**

**Operations), Industry Vertical, Deployment Type, Organization Size, and Region - Global Forecast to 2023.** [S.l.], p. 151. 2018. (TC 6280).

MAYO CLINIC. Epilepsy surgery. **Mayo Clinic**, 2019. Available at: <https://www.mayoclinic.org/tests-procedures/epilepsy-surgery/about/pac-20393981>. Accessed: 20 August 2019.

MIHAESCU, O. P. FREQUENCY DISTRIBUTION. In: SALKIND, N. J. **Encyclopedia of Research Design.** [S.l.]: SAGE Publications, Inc., 2012. p. 503-507.

NARKHEDE, S. Understanding AUC - ROC Curve. **Towards Data Science**, 2018. Available at: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>. Accessed: 29 November 2019.

NARKHEDE, S. Understanding Confusion Matrix. **Towards Data Science**, 2018. Available at: <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>. Accessed: 29 November 2019.

NIST/SEMATECH. Measures of Skewness and Kurtosis. **e-Handbook of Statistical Methods**, 2012. Available at: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm>. Accessed: 8 November 2019.

NOAHTAR, S. et al. A glossary of terms most commonly used by clinical electroencephalographers and proposal for the report form for the EEG findings. **Electroencephalography and Clinical Neurophysiology Supplement**, v. 52, p. 21-41, 1999.

NUNES, M. L. et al. Incidence of epilepsy and seizure disorders in childhood and association with social determinants: a birth cohort study. **Jornal de Pediatria**, Rio de Janeiro, v. 87, n. 1, p. 50-56, January 2011. ISSN 0021-7557.



OLEJNIK, S. OMEGA SQUARED. In: SALKIND, N. J. **Encyclopedia of Research Design**. [S.l.]: [s.n.], 2012. p. 963-967.

PACKT. Expanding Your Data Mining Toolbox. **Packt**, 2016. Available at: <<https://hub.packtpub.com/expanding-your-data-mining-toolbox/>>. Accessed: 21 September 2019.

PAHO. **Epilepsy in Latin America**. Pan American Health Organization (PAHO). Santiago, Chile, p. 108. 2013. (01033214).

PASSARO, E. A.; BENBADIS, S. R. Outcome of Epilepsy Surgery. **Medscape**, 2018. Available at: <<https://emedicine.medscape.com/article/1185416-overview#showall>>. Accessed: 15 August 2019.

PENTAHO. Dimensionality Reduction using Principal Component Analysis. **TenthPlanet**, 2019. Available at: <<https://blog.tenthplanet.in/principle-component-analysis/>>. Accessed: 30 October 2019.

PETR KLIMES et al. NREM sleep is the state of vigilance that best identifies the epileptogenic zone in the interictal electroencephalogram. **Epilepsia**, n. 00, p. 1-12, October 2019.

PIASTA, S. B.; JUSTICE, L. M. COHEN'S D STATISTIC. In: SALKIND, N. J. **Encyclopedia of Research Design**. [S.l.]: SAGE Publications, Inc, 2012. p. 205-210.

PINGEL, J. Ensemble Learning - Deep Learning - MATLAB & Simulink. **Mathworks Blog**, 2019. Available at: <<https://blogs.mathworks.com/deep-learning/2019/06/03/ensemble-learning/>>. Accessed: 14 November 2019.

PRATT, W. E. Wilcoxon Rank Sum Test. In: SALKIND, N. J. **Encyclopedia of Research Design**. [S.l.]: SAGE Publications, Inc. , 2012. p. 1629-1633.

ROSENOW, F.; LÜDERS, H. O. Presurgical evaluation of epilepsy. **Brain: a journal of neurology**, v. 124, n. Pt. 9, 2001. ISSN 1683–700.

RYVLIN, P.; RHEIMS, S. Epilepsy surgery: eligibility criteria and presurgical evaluation. **Dialogues in Clinical Neuroscience**, v. 10, n. 1, p. 91-103, March 2008.

SAS. Introduction to SEMMA. **SAS Help Center**, 2017. Available at: <<https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbj1a2.htm&docsetVersion=14.3&locale=en>>. Accessed: 20 October 2019.

SAZGAR, M.; YOUNG, M. G. EEG Artifacts. In: SAZGAR, M.; YOUNG, M. G. **Absolute Epilepsy and EEG Rotation Review**. [S.l.]: Springer, Cham, 2019. p. 149-162.

SHAFIQUE, U.; QAISER, H. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). **International Journal of Innovation and Scientific Research**, v. 12, n. 1, p. 217-222, November 2014. ISSN 2351-8014.

SHAMSAEI, G. R. Epileptic discharges. In: SHAMSAEI, G. R. **Review Of Clinical Electroencephalography**. [S.l.]: Salekan, 2014. Cap. 8, p. 85-95.

SHARMA, A. Confusion Matrix in Machine Learning. **GeeksforGeeks**, 2019. Available at: <<https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>>. Accessed: 29 November 2019.

SIGMA PLUS STATISTIEK. Pearson Correlations – Quick Introduction. **Sigma Plus Statistiek Tutorials**, 2019. Available at: <<https://www.spss-tutorials.com/pearson-correlation-coefficient/>>. Accessed: 29 November 2019.

SIQUEIRA, H. H. et al. Prevalence of epilepsy in a Brazilian Semiurban Region: An epidemiological study. **Revista Brasileira de Neurologia e Psiquiatria**, v. 20, n. 2, p. 124-138, 2016. ISSN 14140365.

SIRVEN, J. I.; SHAFER, P. O. Challenges with Epilepsy | Epilepsy Foundation. **Epilepsy Foundation**, 2014. Available at: <<https://www.epilepsy.com/learn/challenges-epilepsy>>. Accessed: 12 August 2019.

SIRVEN, J. I.; SHAFER, P. O. Refractory Epilepsy (Difficult to Treat Seizures). **Epilepsy Foundation**, 2014. Available at: <<https://www.epilepsy.com/learn/refractory-epilepsy-difficult-treat-seizures>>. Accessed: 13 August 2019.

STAFSTROM, C. E.; CARMANT, L. Seizures and Epilepsy: An Overview for Neuroscientists. **Cold Spring Harbor Perspectives in Medicine**, 5, n. 6, June 2015. 1-18.

SULLIVAN, L.; LAMORTE, W. W. Outliers and Tukey Fences. **Descriptive Statistics - Boston University School of Public Health**, 2016. Available at: <[http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704\\_summarizingdata/bs704\\_summarizingdata7.html#headingtaglink\\_3](http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_summarizingdata/bs704_summarizingdata7.html#headingtaglink_3)>. Accessed: 29 November 2019.

SYNAPSE. Epilepsy & other seizures. **SYNAPSE - Australia's Brain Injury Organization**, 2019. Available at: <<https://synapse.org.au/information-services/epilepsy-other-seizures.aspx>>. Accessed: 22 May 2019.

TANG, F. et al. Drug-resistant epilepsy: Multiple hypotheses, few answers. **Frontiers in Neurology**, v. 8, n. JUL, p. 301, July 2017. ISSN 1664-2295.

TEPLAN, M. Fundamentals of EEG Measurement. **Measurement Science Review**, v. 2, n. 2, January 2002.

THE MCGILL PHYSIOLOGY VIRTUAL LABORATORY. Biomedical Signals Acquisition: EEG > introduction. **The McGill Physiology Virtual Laboratory**, 2005. Available at: <[https://www.medicine.mcgill.ca/physio/vlab/biomed\\_signals/eeg\\_n.htm](https://www.medicine.mcgill.ca/physio/vlab/biomed_signals/eeg_n.htm)>. Accessed: 25 September 2019.

TREVINO, A. Introduction to K-means Clustering. **Oracle Data Science Blog**, 2016.

Available at: <<https://blogs.oracle.com/datascience/introduction-to-k-means-clustering>>.

Accessed: 29 nov. 2019.

VALLAT, R. Compute the average bandpower of an EEG signal. **Raphael Vallat**, 2018.

Available at: <<https://raphaelvallat.com/bandpower.html>>. Accessed: 1 October 2019.

WHO. Fact Sheets: Epilepsy. **World Health Organization (WHO)**, 7 February 2019.

Available at: <<https://www.who.int/news-room/fact-sheets/detail/epilepsy>>. Accessed: 20

May 2019.

WIESER, H. et al. ILAE Commission Report. Proposal for a new classification of outcome with respect to epileptic seizures following epilepsy surgery. **Epilepsia**, v. 42, n. 2, p. 282-286, February 2001.

WILD, C. The Wilcoxon Rank-Sum Test. **University of Auckland**, 1997. Available at:

<<https://www.stat.auckland.ac.nz/~wild/ChanceEnc/Ch10.wilcoxon.pdf>>. Accessed: 20

October 2019.

WILKE, C. et al. Graph analysis of epileptogenic networks in human partial epilepsy.

**Epilepsia**, v. 52, n. 1, p. 84-93, January 2011.

WITTE, H. et al. Use of discrete Hilbert transformation for automatic spike mapping: A

methodological investigation. **Medical & Biological Engineering & Computing**, v. 29, n. 3,

p. 242-248, June 1991.