UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
PROJETO DE GRADUAÇÃO

ALEXANDRE ROSSETO LEMOS

ANÁLISE DE MICROESTADOS EM SINAIS DE ELETROENCEFALOGRAMA: UM ESTUDO PARA PREVISÃO DE CRISES EM PACIENTES EPILÉPTICOS

VITÓRIA
2020

ALEXANDRE ROSSETO LEMOS

ANÁLISE DE MICROESTADOS EM SINAIS DE ELETROENCEFALOGRAMA: UM ESTUDO PARA PREVISÃO DE CRISES EM PACIENTES EPILÉPTICOS

Parte manuscrita do Projeto de Graduação do aluno **Alexandre Rosseto Lemos**, apresentado ao Departamento de Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Engenheiro Eletricista.

Orientador: Prof. Dr. Patrick Marques Ciarelli

VITÓRIA
2020

ALEXANDRE ROSSETO LEMOS

ANÁLISE DE MICROESTADOS EM SINAIS DE ELETROENCEFALOGRAMA: UM ESTUDO PARA PREVISÃO DE CRISES EM PACIENTES EPILÉPTICOS

Parte manuscrita do Projeto de Graduação do aluno **Alexandre Rosseto Lemos**, apresentado ao Departamento de Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Engenheiro Eletricista.

Aprovada em 17 de dezembro de 2020.

COMISSÃO EXAMINADORA:

Prof. Dr. Patrick Marques Ciarelli
Universidade Federal do Espírito Santo
Orientador

Prof. Dr. Klaus Fabian Côco
Universidade Federal do Espírito Santo
Examinador

MSc. Lorraine Marques Alves
Universidade Federal do Espírito Santo
Examinadora

Dedico este trabalho a todos os meus familiares e amigos, cujo apoio durante estes anos foi fundamental para eu chegar até aqui. Dedico também à minha namorada Jade cujo amor e companhia eu tenho o prazer de ter na minha vida.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pela oportunidade de ter chegado a este momento em minha vida. Agradeço aos meus pais e irmãos, pelos ensinamentos e amor incondicional. Agradeço à minha namorada Jade por sempre me fazer ser uma pessoa melhor e sempre estar ao meu lado para me apoiar. Agradeço aos amigos que sempre me proporcionaram bons momentos ao longo dessa jornada. Agradeço também ao meu orientador Prof. Dr. Patrick Marques Ciarelli pela atenção e paciência, que mesmo tendo muitos compromissos, sempre esteve à disposição quando requisitado Agradeço ainda aos professores e funcionários do Departamento de Engenharia Elétrica da Universidade Federal do Espírito Santo, os quais serei eternamente grato.

RESUMO

Avanços tecnológicos possuem um papel importante no tratamento, prevenção e melhoria de vida de pessoas com doenças neurológicas. A epilepsia é uma das doenças neurológicas mais comuns no mundo, que causa um grande impacto negativo na vida de pessoas que possuem esta doença devido aos ataques epiléticos recorrentes. A detecção de convulsões com antecedência pode dar tempo ao paciente para que ele possa tomar medidas, tais como medicamentos anticonvulsivantes ou se colocando em um ambiente seguro, com o intuito de diminuir os efeitos colaterais causados pela crise. Diversas abordagens em aprendizado de máquinas utilizando sinais de eletroencefalograma (EEG) estão sendo estudadas para realização de algoritmos capazes de predizer as crises com melhores desempenhos. Uma abordagem que foi pouco explorada, mas que possui grande potencial de gerar bons resultados nos estudos da epilepsia, é a análise de microestados. Nesta abordagem, o sinal de EEG é dividido em intervalos quase estáveis, sendo que cada intervalo é associado a uma topografia do sinal de EEG. Neste trabalho, um estudo é proposto para analisar o uso de microestados para a predição de crises em pacientes com epilepsia. Foram estudadas técnicas de análise de dados e aprendizado de máquina para extrair informações relevantes. Observou-se que o melhor resultado de classificação foi obtido quando utilizou-se 16 microestados e o algoritmo máquina de vetores de suporte (SVM, do inglês *support vector machine)* utilizando o *kernel* função de base radial gaussiana (RBF, do inglês *radial basis function*) e com a técnica de seleção de atributos análise de componentes vizinhos (NCA, do inglês *neighborhood component analysis*), atingindo uma acurácia de 92,37%, 92,15% de sensibilidade e 92,58% de especificidade. Também foi possível prever quando uma crise convulsiva iria ocorrer com mais de 12 minutos de antecedência, usando as mesmas técnicas. Essa proposta pode contribuir de modo significativo para obtenção de algoritmos que possam ser utilizados em pacientes epilépticos, tendo potencial também para colaborar em abordagens que aplicam microestados para analisar outras doenças neurológicas.

Palavras-chave: Epilepsia. Predição. Eletroencefalograma. Processamento de sinais. Aprendizado de máquinas. Análise de microestados.

ABSTRACT

Technological advances have an important role in the treatment, prevention and improvement of life of people with neurological diseases. Epilepsy is one of the most common neurological diseases in the world, which has a major negative impact on the lives of people who have this disease due to recurrent epileptic seizures. The detection of seizures in advance can give the patient time to take measures, such as anticonvulsant medications or placing themselves in a safe environment, in order to reduce the side effects caused by the crisis. Several approaches in machine learning using electroencephalogram (EEG) signals are being studied for the realization of algorithms capable of predicting crises with better performances. An approach that has been little explored, but which has great potential to generate good results in studies of epilepsy, is the analysis of microstates. In this approach, the EEG signal is divided into almost stable intervals, each interval being associated with a topography of the EEG signal. In this work, a study is proposed to analyze the use of microstates to predict seizures in patients with epilepsy. Data analysis and machine learning techniques were studied to extract relevant information. It was observed that the best classification result was obtained when 16 microstates and the support vector machine (SVM) algorithm using the kernel radial basis function (RBF) were used with the neighborhood component analysis (NCA) attribute selection technique, reaching an accuracy of 92.37%, 92.15% sensitivity and 92.58 % specificity. It was also possible to predict when a seizure would occur with more than 12 minutes in advance, using the same techniques. This proposal can contribute significantly to obtaining algorithms that can be used in epileptic patients, having the potential also to collaborate in approaches that apply microstates to analyze other neurological diseases.

Keywords: Epilepsy. Prediction. Electroencephalogram. Signal processing. Machine learning. Analysis of microstates.

LISTA DE FIGURAS

Figura 1 – Exemplo de cinco microestados obtidos utilizando o EEGLAB 14
Figura 2 – Exemplo de um classificador do modelo kNN 20
Figura 3 – Alteração de dimensionalidade dos dados de (a) classes não linearmente separáveis para (b) classes separáveis por um hiperplano, realizada pelo algoritmo SVM 21
Figura 4 – Representação de um SVM e seu hiperplano ótimo separando amostras de duas classes 22
Figura 5 – Validação cruzada *hold-out* sendo aplicada a um conjunto de dados 23
Figura 6 – Validação cruzada *k-fold* aplicada a um conjunto de dados 23
Figura 7 – Validação cruzada *leave-one-out* aplicada a um conjunto de dados 24
Figura 8 – EEG realizado em um indivíduo com uma rede contendo 128 eletrodos 26
Figura 9 – Exemplos de sinais de EEG para os períodos (a) interictal e (b) preictal de um canal 27
Figura 10 – Processo de segmentação de microestados: captação do sinal (a), cálculo do GFP (b), obtenção das topografias (c), determinação dos microestados (d) 31
Figura 11 – Canais utilizados na análise de microestados 34
Figura 12 – Exemplo de uma série temporal dividida em 8 segmentos 35
Figura 13 – Microestados criados para análise do número ideal de microestados 36
Figura 14 – Sinal contínuo de EEG 38
Figura 15 – Representação gráfica da quantidade de vezes que cada atributo apareceu no resultado do NCA 40
Figura 16 – Sinal dos pacientes 1 (a) e 2 (b) segmentados em janelas de 40 segundos e classificados pelo SVM RBF 43

LISTA DE QUADROS

Quadro 1 – Resultados de estudos utilizando análises de microestados de sinais de EEG 15
Quadro 2 – Pacientes do *Children's Hospital Boston* analisados neste trabalho.....33
Quadro 3 – Resultados dos modelos utilizando atributos de 4 microestados.....39
Quadro 4 – Resultados dos modelos utilizando atributos de 8 microestados.....39
Quadro 5 – Resultados dos modelos utilizando atributos de 16 microestados.....40
Quadro 6 – Representação dos atributos e seus respectivos microestados.....41
Quadro 7 – Resultados dos modelos utilizando 16 microestados e os atributos selecionados pelo NCA 41
Quadro 8 – Resultados das predições para os 20 pacientes.....42
Quadro 9 – Comparação de resultados obtidos por Raj e outros (2018) e o melhor classificador deste trabalho.....44

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
EEG	Eletroencefalograma
GEV	*Global Explained Variation*
GFP	*Global Field Power*
GUI	*Guided User Interface*
GMD	*Global Map Dissimilarity*
kNN	*k-Nearest Neighbors*
LOO	*Leave-One-Out*
NCA	*Neighborhood Component Analysis*
RBF	*Radial Basis Function*
SVM	*Support Vector Machine*
UFES	Universidade Federal do Espírito Santo

LISTA DE SÍMBOLOS

D_E	Distância Euclidiana entre dois pontos
x_i	Ponto no espaço no instante i
xj	Ponto no espaço no instante j
d	Dimensões do espaço
D_M	Distância Mahalanobis entre dois objetos
$Corr_k$	Correlação espacial média do k-ézimo microestado
a_{li}	Protótipo de microestado no instante i
V	Potencial instantâneo no eletrodo (μV)
C	Quantidade de eletrodos
t	Tempo (s)
Dur	Duração média (ms)
WT	Tempo total do período analisado do sinal (ms)
MST	Período do sinal onde existe a presença do microestado analisado (ms)
CON	Quantidade de vezes seguidas em que o microestado analisado aparece no sinal
TC	Tempo de cobertura (%)
AMP	Picos de potência do campo global (μV)
GFPP	Quantidade de microestados de uma classe analisada
N_k	Quantidade de amostras rotuladas ao k-ézimo microestado
a_k	Mapa protótipo para o *k*-ézimo microestado

SUMÁRIO

1 INTRODUÇÃO **12**
1.1 Apresentação e Objetivo da Pesquisa **12**
1.2 Justificativa **14**
1.3 Objetivos **17**
1.3.1 Objetivo Geral 17
1.3.2 Objetivos Específicos 17
1.4 Organização do Trabalho **17**
2 EMBASAMENTO TEÓRICO **19**
2.1 Algoritmo k Vizinhos Mais Próximos **19**
2.2 Máquina de Vetores de Suporte **20**
2.3 Métodos de Validação Cruzada **22**
2.3.1 Validação *Hold-Out* 22
2.3.2 Validação Cruzada *k-Fold* 23
2.3.3 Validação Cruzada *Leave-One-Out* 24
2.4 Análise de Componentes Vizinhos **25**
2.5 Eletroencefalograma **26**
2.6 Análise de Microestados **27**
2.7 *Backfit* **30**
3 METODOLOGIA **32**
3.1 Descrição do Banco de Dados **32**
3.2 Pré-Processamento **34**
3.3 Análise em Microestados **35**
3.4 Extração e Seleção de Características **36**
3.5 Modelos de Classificação **37**
3.6 Método de Previsão da Crise Epiléptica **37**
4 RESULTADOS **39**
5 CONCLUSÕES **45**
REFERÊNCIAS BIBLIOGRÁFICAS **47**

1 INTRODUÇÃO

1.1 Apresentação e Objetivo da Pesquisa

A epilepsia é um distúrbio neurológico caracterizado por gerar ataques epilépticos recorrentes nos indivíduos afetados, resultando em diversos efeitos cognitivos, psicológicos e consequências sociais. Um ataque epilético é descrito como uma mudança comportamental passageira que pode apresentar sinais objetivos (convulsão) ou sintomas subjetivos (como perda de consciência e enrijecimento dos membros), causada por alterações na atividade neural do cérebro (DEVINSKY et al., 2018).

Estima-se que existam cerca de 50 milhões de pessoas ao redor do mundo que sofram com a epilepsia, sendo aproximadamente 80% cidadãos de países subdesenvolvidos ou em desenvolvimento, e cerca de 5 milhões de novos casos são diagnosticados a cada ano (WORLD HEALTH ORGANIZATION, 2020). Os pacientes sofrem constantemente por conta da falta de informação que o resto da população tem sobre a doença, fazendo com que portadores da epilepsia e seus familiares sejam alvos de discriminação, impactando diretamente em suas qualidades de vida. Além do efeito social, pessoas afetadas pela epilepsia tendem a ter mais problemas físicos (causados pelas convulsões), psicológicos (como ansiedade e depressão) e possuem o triplo da taxa de mortalidade prematura se comparado ao resto da população (WORLD HEALTH ORGANIZATION, 2020).

Os avanços tecnológicos nas áreas de aprendizado de máquina estão ajudando o homem a conseguir detectar, com antecedência, quando crises epilépticas irão ocorrer. Um dos grandes problemas desta doença é que a perda de consciência e súbita perda do controle motor geralmente ocorrem sem aviso. Com isso, a capacidade de prever crises epiléticas ajudaria a tranquilizar e reduzir a ansiedade dos pacientes, melhorando assim a qualidade de vida. Como atualmente não é possível controlar completamente os ataques epilépticos de um paciente, a previsão de crises se torna uma alternativa importante do manejo clínico e tratamento (MORMANN et al., 2007).

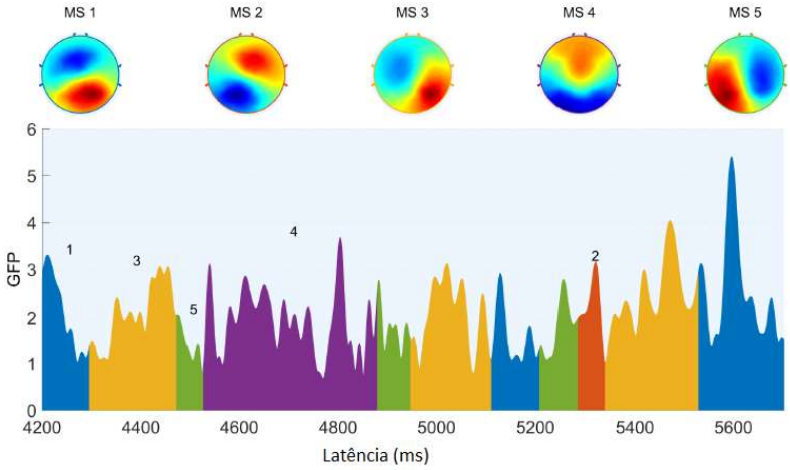
Um método bem estabelecido para caracterizar as atividades do cérebro humano usando eletroencefalografia (EEG) multicanal, e que está ganhando muita repercussão, é o de análise

de microestados. Ele envolve o exame de microestados elétricos no cérebro, definidos como curtos períodos sucessivos durante os quais a configuração do campo potencial do couro cabeludo permanece semi-estável, sugerindo quase simultaneidade de atividade entre os nós da rede neural cerebral, ou seja, entre os neurônios do cérebro. Alguns protótipos de microestados, recorrentes ao longo do tempo, podem ser identificados de forma confiável entre diversos pacientes, indicando a possibilidade de identificação de padrões (MICHEL; KOENIG, 2018).

A análise de microestados de EEG tornou-se uma popular ferramenta para caracterizar a dinâmica espaço-temporal de dados eletrofisiológicos em larga escala. Foi bastante usada em vários estudos de EEG e os microestados descobertos foram associados a diversas funções e doenças cerebrais. Presume-se que os microestados de EEG apresentam a característica de que, em um determinado momento, a topografia está em um estado apenas e possuam transição discreta de um estado para outro (MISHRA; ENGLITZ; COHEN, 2019).

Uma ferramenta computacional para a obtenção e análise de microestados de sinais de EEG é o EEGLAB (POULSEN et al., 2018). Essa *toolbox*, desenvolvida para o *Matlab* e com código aberto, foi projetada especificamente para facilitar o desenvolvimento de novos métodos. A Figura 1, obtida através da *toolbox*, mostra um sinal de EEG após ser rotulado em 5 microestados. Este mapeamento ocorre durante o processo de *backfit,* uma das etapas de processamento dos sinais. Cada faixa de cor do sinal é associada a um microestado, conforme pode ser observada na Figura 1. Embora o EEGLAB possa ser controlado com uma interface gráfica do usuário (GUI, do inglês *guided user interface*), o *Matlab* permite aos usuários criarem seus próprios scripts para modificar a análise conforme desejarem. Este trabalho tem como objetivo propor uma nova abordagem de análise, com base em mudanças no código executado, possibilitando assim uma simulação de um processamento *on-line* dos dados, uma vez que o algoritmo original não possui essa funcionalidade. Isso irá proporcionar veracidade ao modelo da predição de crises epilépticas.

Figura 1 – Exemplo de cinco microestados obtidos utilizando o EEGLAB



Fonte: Poulsen e outros (2018).

1.2 Justificativa

A análise de microestados é uma representação alternativa do sinal de EEG, sendo cada vez mais aplicada e baseada em uma análise topográfica (POULSEN et al., 2018). Existem relações comprovadas em estudos entre as classes de microestados de EEG e alterações psicopatológicas em doenças cerebrais (NISHIDA et al., 2013).

Nishida e outros (2013) destacam a importância da dinâmica cerebral de subsegundos para a compreensão dos distúrbios psiquiátricos, como a demência, esquizofrenia e a doença de Alzheimer. Estudos publicados mostram a existência de certos padrões na distribuição dos microestados, dependendo da doença neurológica analisada (Quadro 1).

Quadro 1 – Resultados de estudos utilizando análises de microestados de sinais de EEG

Estudo	Ano	Doença	Medica-ção	Pacien-tes	Microestados			
					A	B	C	D
König	1999	Esquizofre-nia	Não	9				↓ Dur Topo
Strelets	2003	Esquizofre-nia	Não	14		↓ Dur Topo		
Lehmann	2004	Esquizofre-nia	Não	27	↑Occ ↑Cov	↓Dur	↑Occ	↓Dur
Kikuchi	2007	Esquizofre-nia	Não	21		↑Occ	↑Occ	↓ Dur ↓Cov
Nishida	2013	Esquizofre-nia	Não	21	↓Dur ↑Occ	↓Dur	↑Occ	↓Dur
Andreaou	2013	Esquizofre-nia	Sim	18		↑Cov		
Tomescu	2015	Esquizofre-nia	Sim	27			↑Dur ↑Occ ↑Cov ↑GEV	↓Occ ↓Cov ↓GEV
Irisawa	2006	Esquizofre-nia	Não	24		↓Dur		
Kindler	2011	Halucina-ção	Sim	9				↓Dur
Andreaou	2013	Esquizofre-nia de alto risco	Sim	18	↑Cov ↑Occ			
Tomescu	2014	22q11DS	Sim	30			↑Dur ↑Occ ↑Cov ↑GEV	↓Dur ↓Occ ↓Cov ↓GEV
Nishida	2013	Demência	Não	18			↓Dur	
Kikuchi	2011	Transtorno do pânico	Não	18	↑Cov ↑Occ		↓Occ	
Gschwind	2016	Esclerose múltipla	Não	53	↑Dur	↑Dur		
Corradini	2014	Ferimento na cabeça	Não	26	↓Dur		↓Dur	↓Dur
Drissi	2016	Narcolepsia	Não	16	↓Dur ↓Cov ↓GEV	↑Dur ↑Cov ↑GEV	↑Dur ↑Cov ↑GEV	
Gao	2016	Diplegia espástica	Não	15	↑Occ	↓Dur	↑Occ	
Zappasodi	2017	AVC	Não	47			↓Occ Topo avc esquer-do	↓Cov Topo avc direito

Fonte: Michael e Koenig (2018).
Nota[1]: Adaptado pelo autor.
Nota[2]: Resultados de diversos estudos para diferentes doenças neurológicas utilizando método de agrupamento k-means e 4 microestados.
Nota[3]: Dur = Duração, Occ = Ocorrência, Cov = Cobertura, Topo = Topografia.

Nos últimos anos, as pesquisas sobre a previsão da crise relacionada a pacientes com epilepsia aumentaram por ser um problema mais desafiador, devido à dificuldade de extrair detalhes minuciosos dos sinais do eletroencefalograma (EEG) para antecipar com precisão uma convulsão. Acharya, Hagiwara e Adeli (2018) revisaram diferentes algoritmos de previsão de crises e seus desafios apresentados nas atuais pesquisas. Um dos problemas relacionados a esse estudo é a generalização do algoritmo que possa ser aplicado em todos os pacientes, independentemente do tipo de convulsão. Portanto, ainda se apresenta a necessidade de criar modelos com maior precisão e confiabilidade e que possam ser aplicados em dispositivos portáteis para uso diário dos pacientes (ACHARYA; HAGIWARA; ADELI, 2018). Para tal finalidade, uma metodologia que possa ser implementada de forma *on-line* está mais próximo da realidade esperada pelos pacientes com epilepsia.

Foram identificados poucos trabalhos que usaram a análise de microestados para epilepsia. Recentemente, Raj e outros (2018) elaboraram um estudo exploratório para investigar a possibilidade de distinguir 21 pacientes com epilepsia de 21 pessoas saudáveis, por meio de análise de microestados. Parâmetros como a duração média, frequência de ocorrência e a porcentagem de cobertura de tempo de 4 classes de microestados foram utilizados para realização da classificação. Para avaliação da utilidade dos microestados em registros de EEG epiléptico, dois algoritmos de aprendizado de máquina foram empregados, a saber: a análise discriminante linear de Fisher, que resultou em uma acurácia geral de 73,80%, com sensibilidade de 81,0% e especificidade de 66,7%, e a regressão logística, que obteve, para as mesmas métricas, 66,7%, 71,4% e 61,9%, respectivamente (RAJ et al., 2018).

Piorecka e outros (2018) também demonstraram que a análise do microestado é viável para os registros de EEG de pacientes com atividade epiléptica. Eles realizaram um estudo com 10 pessoas saudáveis e 6 diagnosticadas com epilepsia para distinguir entre gravações de EEG epilépticas e não epilépticas usando a análise de microestados. Os parâmetros duração, ocorrência e cobertura foram calculados a partir de 4 classes de microestados, entretanto, não foram utilizados métodos de classificação para avaliar o desempenho do método. A pesquisa contribuiu para comprovar que a dinâmica dos microestados é diferente entre o grupo de controle e os pacientes com epilepsia, possibilitando maiores estudos sobre a aplicação de microestados na epilepsia e a implementação de aprendizado de máquina usando microestados, visto que é uma abordagem nova para essa área de estudo (PIORECKA et al., 2018).

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo geral deste trabalho de graduação foi desenvolver um algoritmo para a predição de crises em pacientes com epilepsia, baseado em algoritmos de classificação, a partir de microestados do sinal de EEG. A análise foi realizada empregando-se abordagens estatísticas e de aprendizado de máquina.

1.3.2 Objetivos Específicos

- Obter um banco de dados público de sinais de EEG de pacientes com epilepsia;
- Alterar o algoritmo de Poulsen e outros (2018) para permitir a realização de análises *on-line* de sinais de EEG por meio de microestados;
- Avaliar a qualidade das análises realizadas por meio de métricas adequadas.

1.4 Organização do Trabalho

Este projeto está dividido em 5 seções, cada um com o conteúdo exposto a seguir.

A seção 1 apresenta uma introdução aos assuntos abordados ao longo do trabalho. São evidenciados estudos sobre epilepsia, análise de microestados e como a utilização de técnicas de aprendizagem de máquina está sendo empregada para entender e diagnosticar melhor doenças neurológicas. Também são expostos os objetivos almejados durante a realização deste trabalho.

Na seção 2 são apresentados os embasamentos teóricos das técnicas empregadas durante o desenvolvimento do projeto. Nele são explicados o funcionamento de diferentes classificadores utilizados, métodos de validação e processamento de dados, bem como o que são sinais de eletroencefalograma e suas particularidades quando se trata de pacientes que sofrem com epilepsia, e como é feito o procedimento de análise de microestados.

A seção 3 expõe qual foi a metodologia empregada durante o desenvolvimento deste projeto, descrevendo o banco de dados utilizado, o pré-processamento dos dados, como foi feita a separação dos sinais em janelas amostrais e a análise de microestados aplicada. Também estão presentes a forma como foram extraídas as características utilizadas e as técnicas de validação cruzada e otimização de atributos adotadas durante o desenvolvimento dos classificadores.

A seção 4 exibe os resultados obtidos com a metodologia apresentada na seção 3, utilizando 4, 8 e 16 microestados. Com isso, torna-se possível uma comparação dos diferentes cenários desenvolvidos.

A seção 5, por fim, é apresentada a conclusão do projeto seguido pelas observações finais. Nela são apresentadas considerações a respeito das diversas análises realizadas durante o desenvolvimento. Também são expostas sugestões para futuros estudos na área que podem auxiliar na obtenção de melhores resultados.

2 EMBASAMENTO TEÓRICO

Esta seção irá apresentar a base teórica utilizada na elaboração do trabalho. Nela, serão expostas as definições gerais dos algoritmos k vizinhos mais próximos, máquina de vetores de suporte, métodos de validação cruzada *hold-out*, *k-fold* e *leave-one-out* (LOO), o método de análise de componentes para seleção de atributos, eletroencefalograma, análise de microestados e como é feito o mapeamento que faz um sinal de EEG ser representado por microestados, chamado de *backfit*.

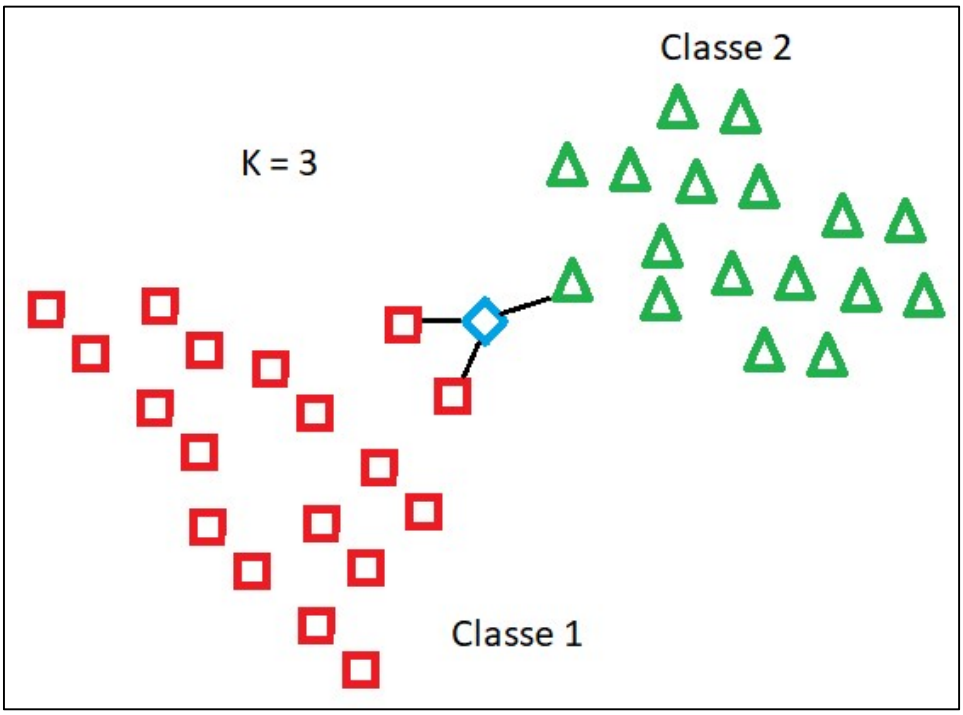
2.1 Algoritmo k Vizinhos Mais Próximos

Dentre as diversas técnicas de aprendizado de máquinas presentes na literatura, o algoritmo k vizinhos mais próximos (kNN, do inglês *k-nearest neighbors*) se caracteriza por ser um método simples porém efetivo para ser utilizado como classificador. Ele é um modelo de aprendizado supervisionado baseado em cálculos de distância e que possui como único parâmetro para ser otimizado o valor de vizinhos mais próximos (*k*). Além disso, é considerado um algoritmo "preguiçoso" (*lazy learning*), pois o aprendizado não ocorre até a etapa de classificação (GUO et al., 2003).

O kNN faz a classificação de um novo dado usando os dados vizinhos a ele. Para cada amostra sem rótulo a ser classificada, encontra-se quais são os *k* elementos mais próximos à ela usando uma métrica de distância como a Euclidiana. A equação (1) ilustra como é calculada a distância Euclidiana (D_E) entre pontos (x_i e x_j) com *d* dimensões. Em seguida, verifica-se, entre os *k* vizinhos, qual é a classe mais comum entre eles e classifica-se o dado analisado como pertencente a esse conjunto Na Figura 2, tem-se um exemplo onde o algoritmo classifica uma amostra (losango azul) como pertencente à uma das duas classes (quadrados vermelhos ou triângulos verdes). Definindo $k = 3$, encontram-se os 3 elementos cujas distâncias (linhas pretas) são as menores da amostra. O classificador irá rotular o dado novo como pertencente à classe mais comum entre os *k* vizinhos mais próximos, no caso, a classe 1. Como o resultado é obtido através de um sistema de votação, é boa prática utilizar valores ímpares na hora de selecionar o número de *k*, a fim de evitar um empate (XIAO, 2018).

$$D_E(x_i, x_j) = \sqrt{\sum_{m=1}^{d} \left(x_{j,m} - x_{i,m}\right)^2} \tag{1}$$

Figura 2 – Exemplo de um classificador do modelo kNN



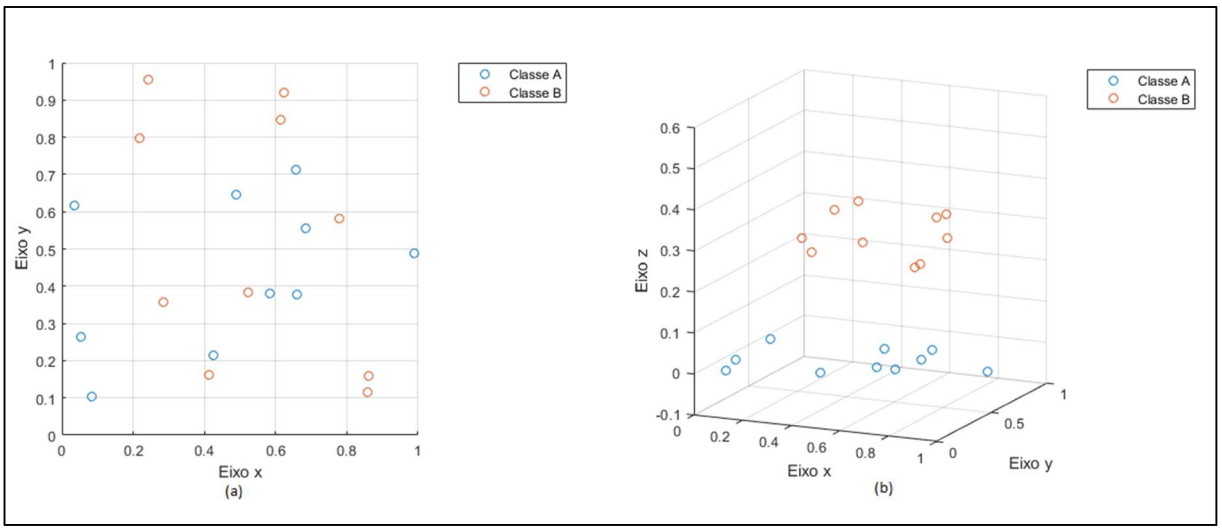
Fonte: Produzido pelo próprio autor.

2.2 Máquina de Vetores de Suporte

Máquina de vetores de suporte (SVM, do inglês *support vector machine)* tem demonstrado ser uma ferramenta eficaz para atuar em problemas de classificação. Seu funcionamento transforma os dados de entrada, geralmente elevando sua dimensionalidade, para possibilitar que eles sejam linearmente separáveis (CORTES; VAPNIK, 1995).

Na Figura 3 é possível visualizar como um modelo SVM consegue obter um hiperplano que separe linearmente os dados de duas classes. Em um plano 2-D as classes não são linearmente separáveis (Figura 3(a)), ou seja, não é possível traçar uma reta que separe completamente os dados da classe A dos dados da classe B. A utilização de um *kernel* para aumentar a dimensionalidade, levando os dados para um plano 3-D, torna possível determinar um hiperplano que realize a separação das classes (Figura 3(b)).

Figura 3 – Alteração de dimensionalidade dos dados de (a) classes não linearmente separáveis para (b) classes separáveis por um hiperplano, realizada pelo algoritmo SVM



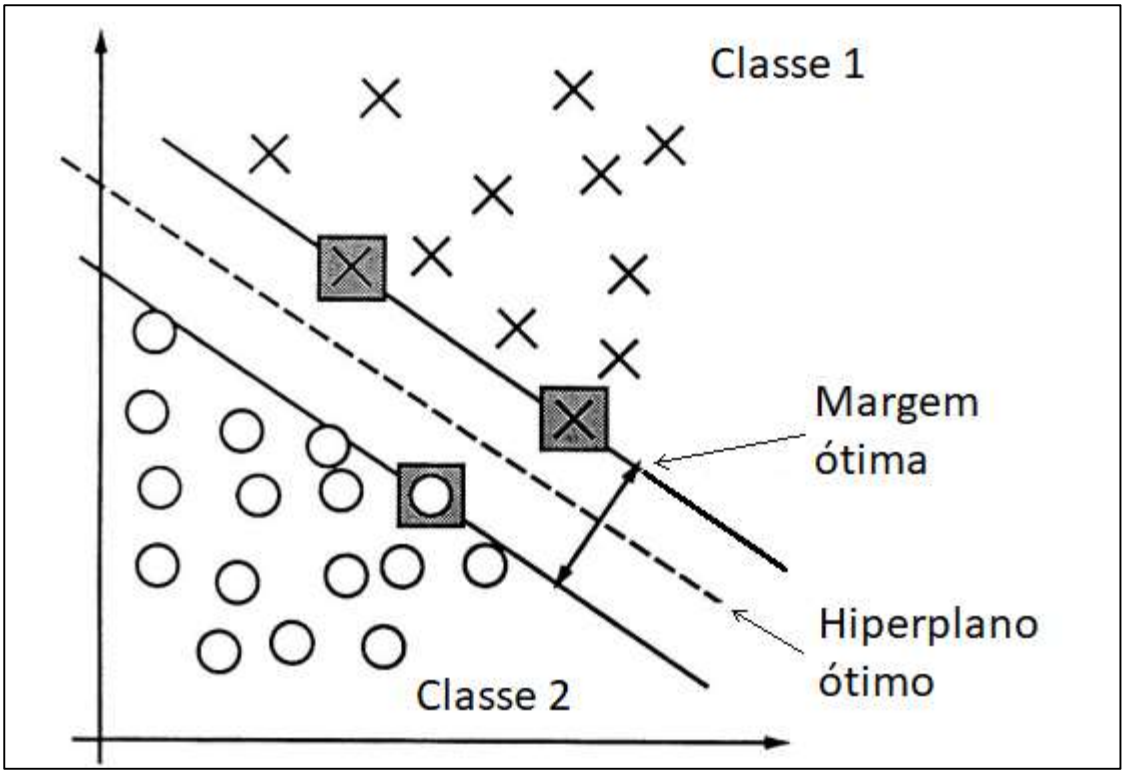
Fonte: Produzido pelo próprio autor.

Uma das principais vantagens do SVM é a sua capacidade de generalização bem robusta, com alta precisão de previsão de classificação. O SVM é formulado como um problema de otimização, onde busca-se encontrar o melhor hiperplano que separe as amostras de cada classe e cuja distância dos vetores de suporte seja a maior possível (margem ótima), conforme ilustrado na Figura 4. Os vetores auxiliares são determinados a partir das amostras mais próximas do plano de decisão (CORTES; VAPNIK, 1995).

Algoritmos SVM usam diferentes conjuntos de funções matemáticas, definidos como o *kernels*, para realizar transformações nos dados de entrada. Existem diversos tipos de *kernels*, sendo os mais comuns o linear, polinomial e função de base radial gaussiana (RBF, do inglês *radial basis function*) (XIAO, 2018).

A capacidade de um SVM predizer dados futuros está relacionada à alta competência de generalização do algoritmo. Quando um novo dado, não utilizado durante a fase de treinamento, é alimentado ao classificador, o modelo consegue classificá-lo mesmo sendo a primeira vez que aquele registro é analisado pelo SVM. Um modelo SVM que utiliza o *kernel* RBF possui dois parâmetros que podem ser otimizados durante a fase de validação, sendo eles γ (gamma) e a função de custo, enquanto que um modelo linear possui apenas a função de custo (OLIVEIRA JUNIOR, 2010).

Figura 4 – Representação de um SVM e seu hiperplano ótimo separando amostras de duas classes



Fonte: Cortes e Vapnik (1995).
Nota: Adaptado pelo autor.

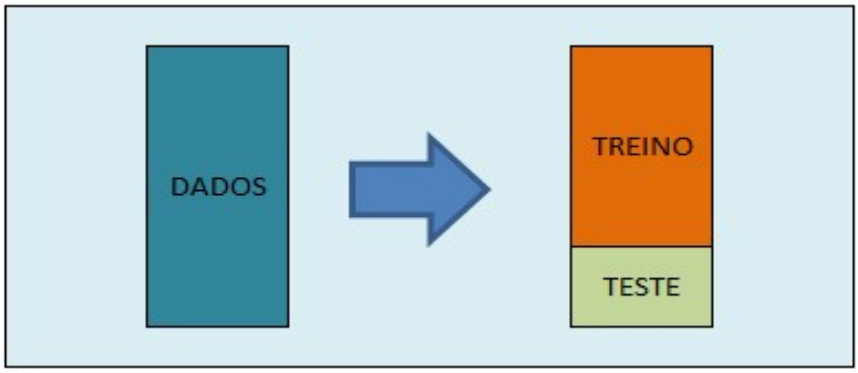
2.3 Métodos de Validação Cruzada

A validação cruzada é um procedimento de validação utilizado como forma de se obter uma estimativa do desempenho geral do modelo, visto que a sua qualidade está diretamente associada a sua capacidade de generalização. Existem diversas técnicas de validação cruzada, que separam os dados em subconjuntos complementares de treino e teste de diferentes formas (MORENO-TORRES; SÁEZ; HERRERA, 2012).

2.3.1 Validação *Hold-Out*

Uma das técnicas de validação cruzada utilizada no desenvolvimento deste projeto foi a *hold-out validation*. Seu procedimento consiste em dividir os dados em conjuntos de treino e de teste, conforme mostrado na Figura 5, utilizando-os separadamente para cada etapa. Desta forma, pode-se avaliar como o modelo irá se comportar quando confrontado com dados nunca antes vistos durante o seu treinamento (KOHAVI, 1995).

Figura 5 – Validação cruzada *hold-out* sendo aplicada a um conjunto de dados

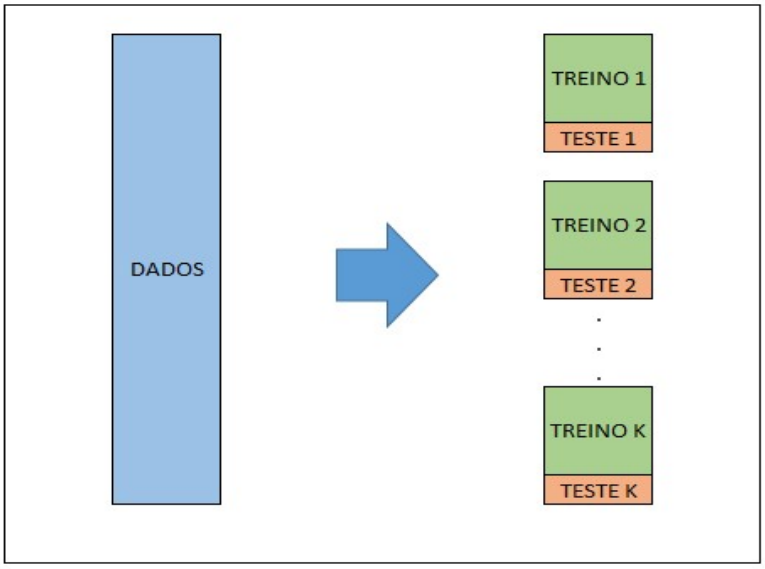


Fonte: Produzido pelo próprio autor.

2.3.2 Validação Cruzada *k-Fold*

O método *k-fold cross-validation* também foi utilizada neste projeto. Essa técnica divide os dados em *k* conjuntos de tamanhos aproximadamente iguais e mutuamente exclusivos, cada um com seu subconjunto de treino e teste. Em seguida, treina-se o modelo *k* vezes, utilizando, em cada iteração, os diferentes dados para treinar e testar o classificador, de acordo com a Figura 6. Ao final do procedimento, o resultado é obtido calculando-se as médias das métricas de desempenho individuais obtidas em cada iteração (KOHAVI, 1995).

Figura 6 – Validação cruzada *k-fold* aplicada a um conjunto de dados

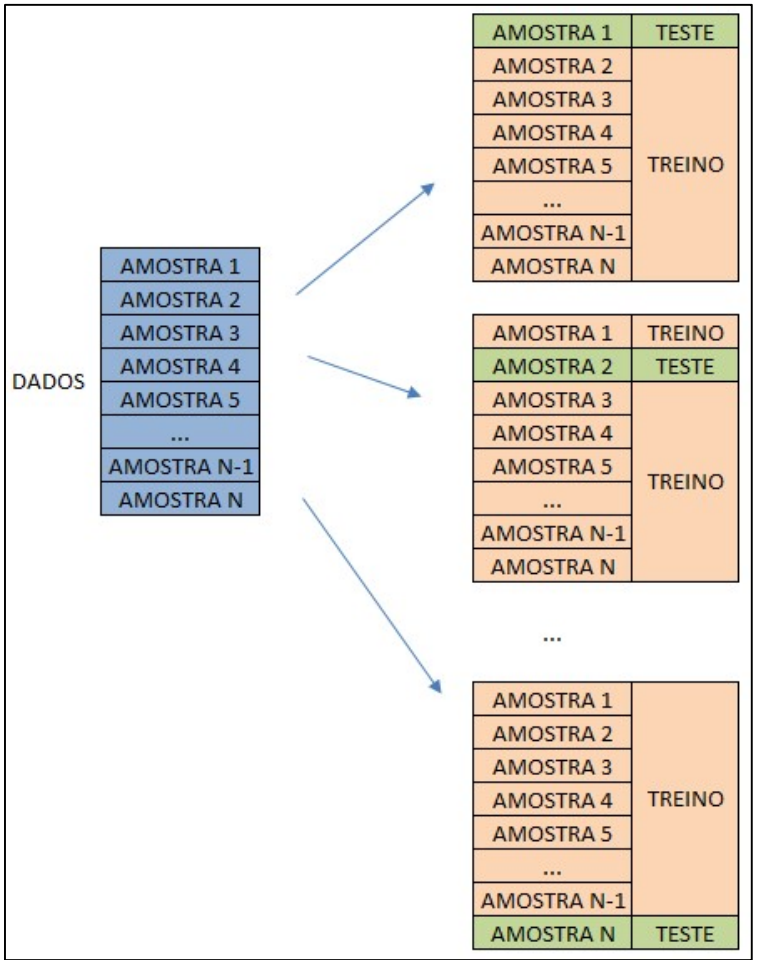


Fonte: Produzido pelo próprio autor.

2.3.3 Validação Cruzada *Leave-One-Out*

A *leave-one-out* (LOO) *cross-validation* foi outra técnica utilizada no desenvolvimento deste projeto. Ela é um caso especial do método *k-fold cross-validation*, quando o valor de *k* é igual ao número total de amostras do banco de dados. Sendo assim, utiliza-se todas as amostras para treino, exceto uma que será utilizada para teste. Esse processo se repete para todas as amostras. Algumas vantagens desse método é que ele não precisa separar os dados de maneira aleatória e nele são utilizados o máximo possível de dados para o treinamento (WONG, 2015). A Figura 7 ilustra a aplicação do LOO em uma base de dados.

Figura 7 – Validação cruzada *leave-one-out* aplicada a um conjunto de dados



Fonte: Produzido pelo próprio autor.

2.4 Análise de Componentes Vizinhos

Durante o desenvolvimento de modelos, especialmente se tratando de dados de alta dimensionalidade, pode-se optar por estudar os atributos utilizados pelos classificadores com o intuito de obter um subconjunto que otimize os resultados obtidos. Uma das técnicas de obtenção desse grupo é a análise de componentes vizinhos. Esse algoritmo é baseado em distância e é muito utilizado em seleções de atributos de dados multidimensionais por sua alta escalabilidade. Ele recebe um vetor de atributos e concede um peso, utilizando um parâmetro de regularização, a cada um deles baseado em resultados obtidos em classificações randomizadas. Ao final, se a característica possuir um peso baixo, significa que ela não agrega informação ao modelo em desenvolvimento (GOLDBERG et al., 2005).

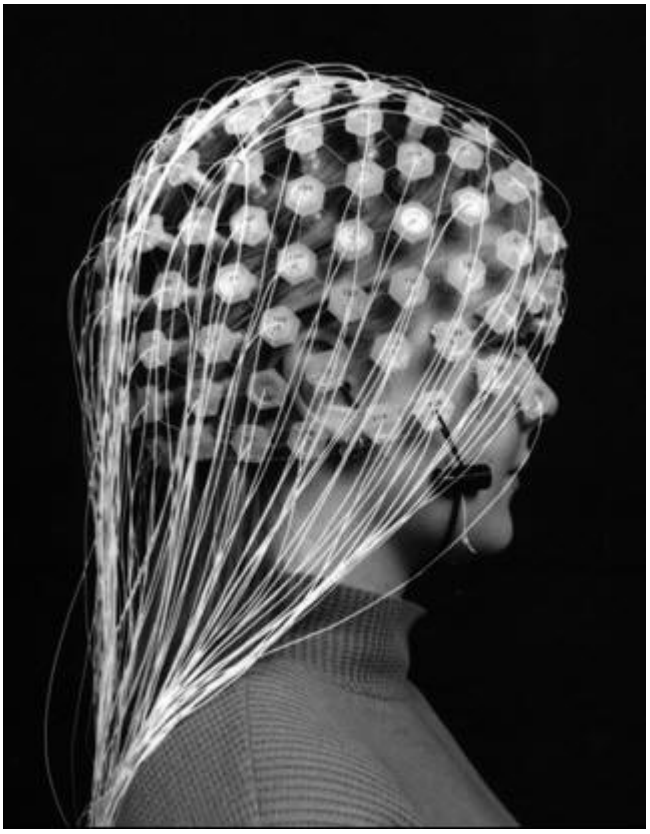
A análise de componentes vizinhos (NCA, do inglês *neighborhood component analysis*) foi desenvolvida para resolver dois problemas presentes no classificador kNN, por consequência disso, as classificações randomizadas realizadas pelo algoritmo são semelhantes ao método do algoritmo de k vizinhos mais próximos. Os desafios a serem resolvidos eram a necessidade de utilizar todos os dados de treinamento para classificar uma nova amostra e qual a melhor métrica de distância a ser utilizada. Para resolver essas questões, foi aplicada uma técnica chamada de *leave-one-out*. Primeiro, classifica-se apenas um dado usando como base os seus *k*-vizinhos mais próximos, utilizando uma medida de distância como métrica, em geral a Mahalanobis (D_M), descrita pela equação (2), onde x_i e x_j são dois pontos e S é a matriz de covariância entre eles. Em seguida, determina-se o raio ótimo de distância para maximizar a acurácia obtida pelo classificador. Por fim, baseado nos resultados obtidos nas classificações, o algoritmo concede um peso ao vetor de atributos, quanto melhor for o resultado, maior o peso, e vice versa. As características que possuírem um peso baixo são consideradas de pouca relevância e podem então ser desconsideradas, com isso, o algoritmo irá selecionar os atributos que mais agregam informações úteis para a classificação (HANOUSEK, 2018).

$$D_M(x_A, x_B) = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)} \tag{2}$$

2.5 Eletroencefalograma

Um eletroencefalograma registra as oscilações dos potenciais elétricos do cérebro através de eletrodos ligados ao couro cabeludo humano, conforme indicado na Figura 8. Existe a possibilidade de se obter os sinais de maneira intracraniana, porém existe uma preferência de conectar os eletrodos no couro cabeludo por ser uma técnica não invasiva. Os sinais de EEG têm origem, em sua maioria, na camada externa do cérebro (córtex cerebral), considerado ser responsável por grande parte dos pensamentos, emoções e comportamento de um indivíduo. O EEG consegue monitorar o estado de consciência de uma pessoa e a análise do sinal da diferença de potencial do couro cabeludo conta uma parte do funcionamento do cérebro. Métodos de análise de sinais de EEG permitem identificações robustas de estágios distintos do sono, profundidade de anestesias e convulsões epilépticas (NUNEZ; SRINIVASAN, 2007).

Figura 8 – EEG realizado em um indivíduo com uma rede contendo 128 eletrodos

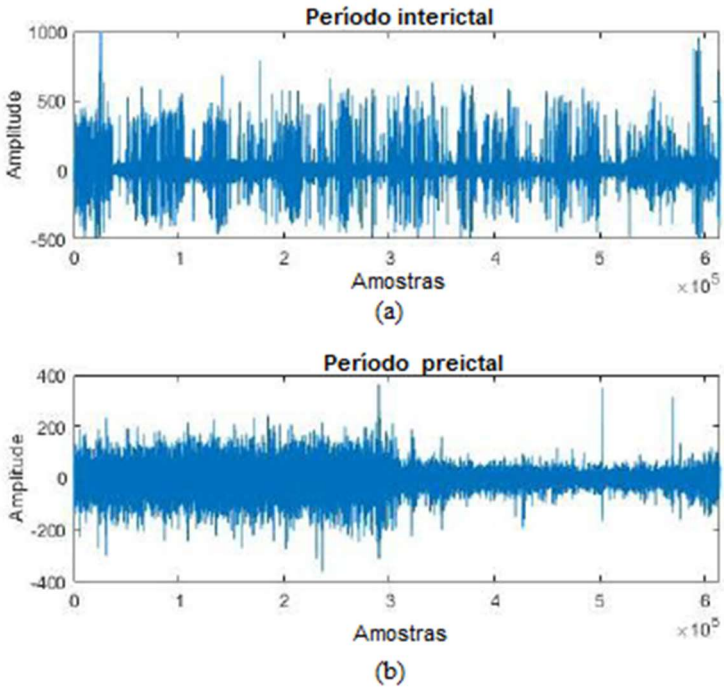


Fonte: Nunez e Srinivasan (2007).

Para o desenvolvimento de modelos de predição e detecção de crises epilépticas, são importantes três estágios de uma crise: o preictal (que a antecede), o ictal (onde ela ocorre) e o

interictal (período entre crises). A detecção de uma crise epiléptica emprega algoritmos que visam detectar o período ictal. Enquanto para previsão da crise geralmente são estudadas as diferenças entre o período de convulsões (estados interictais) e período de transição para o estado ictal, também conhecido como estado preditivo (estado preictal) (KILL et al., 2019). Na Figura 9 são mostrados dois sinais de EEG, um para período interictal e outro para um período preictal.

Figura 9 – Exemplos de sinais de EEG para os períodos (a) interictal e (b) preictal de um canal



Fonte: Kill e outros (2019).
Nota: Adaptado pelo autor.

2.6 Análise de Microestados

O EEG do couro cabeludo corresponde à atividade cerebral superficial e fornece informações sobre o estado geral do cérebro, ao invés de informações localizadas (ASSI et al., 2017). Um determinado eletrodo detectará, além da atividade neural local, outras atividades de fontes potencialmente remotas. Isso ocorre devido a uma correlação intrínseca entre os sinais registrados nesses eletrodos (MICHEL; KOENIG, 2018).

A atividade cerebral observada pelo EEG pode ser descrita por um conjunto de protótipos de topografias de potenciais elétricos do couro cabeludo, em que cada topografia é definida por um período curto de tempo, entre 60 a 120 ms, podendo-se considerar um momento em que a configuração do campo potencial do couro cabeludo permanece estável (MICHEL; KOENIG, 2018). Esses estados discretos estáveis, chamados de microestados, reflete o estado momentâneo da atividade neuronal global, pois essa análise considera simultaneamente o sinal de todos os eletrodos presentes no escalpo (KHANNA et al., 2015, MICHEL; KOENIG, 2018).

Neste trabalho, a obtenção dos microestados foi feita através do *toolbox* EEGLAB, o qual segmenta as gravações individuais de EEG de cada paciente em microestados usando um método de agrupamento, no caso uma versão modificada do k-*means*. O número ideal de *cluster* é calculado para cada conjunto de dados a partir de critérios robustos de otimização (MICHEL; KOENIG, 2018). O número de microestados determinado deve ser o suficiente para que os *clusters* sejam capazes de explicar e se ajustarem aos sinais de EEG (POULSEN et al., 2018).

Para definir o protótipo de microestado mais semelhante a cada amostra, deve ser encontrado o valor da variação global explicada (*GEV*, do inglês *global explained variation*), que é calculada como mostrado na equação (3) (POULSEN et al., 2018).

$$GEV_i = \frac{(Corr\,(x_i, a_{\mathrm{li}})GFP_N)^2}{\sum_{n'}^{N} GFP_{n'}^2} \tag{3}$$

GEV é obtido por meio da correlação ($Corr$) ao quadrado entre a amostra de canais de EEG em um instante (x_{i}) e seu protótipo de microestado (a_{li}), ponderado pela fração da amostra de EEG do total da GFP ao quadrado. Assim, para extrair os microestados, é necessário também calcular o *global field power* (*GFP*), que é a raiz da média das diferenças de potencial ao quadrado de todos os C eletrodos. Na equação (4) é apresentado o cálculo de *GFP*, sendo V_{i} o potencial instantâneo no eletrodo i e V_{mean} a média de potenciais instantâneos nos eletrodos (KHANNA et al., 2015).

$$GFP = \sqrt{((\sum_i^C V_i(t) - V_{mean}(t))^2)/C} \tag{4}$$

O *GFP* mede a resposta global do cérebro a um evento, em que as topografias do campo elétrico nos máximos locais da curva *GFP* são consideradas os microestados do EEG (KHANNA et al., 2015).

Para analisar estatisticamente os microestados, existem métodos validados e bastante utilizados em estudos (MICHEL; KOENIG, 2018). O emprego desses métodos é importante para que se possa extrair informações a partir dos dados. Depois que os *clusters* são determinados, eles são ajustados aos dados de EEG do paciente para definir os microestados. Em seguida, os diferentes parâmetros temporais são obtidos para cada indivíduo e são comparados entre condições experimentais ou entre grupos de participantes (MICHEL; KOENIG, 2018).

As características utilizadas como atributos no modelo do classificador foram obtidas por meio do EEGLAB. Após os sinais serem mapeados por microestados, calcula-se, além do GEV, outros 5 parâmetros para cada microestado, obtidos pelas equações (4) a (7) (FU; CHEN; XIONG, 2018):

- Duração média (milissegundos): período, dentro do tempo total analisado (WT), onde existe a presença do microestado (MST) dividido pela quantidade de vezes seguidas que o microestado aparece (CON), conforme mostrado na equação (5).

$$Dur = \frac{MST}{CON} \tag{5}$$

- Ocorrência (1/s): quantidade de ocorrências do microestado por segundo;
- Tempo de cobertura (%): período onde existe a presença do microestado, dividido pelo tempo total de análise, conforme a equação (6);

$$TC = \frac{MST}{WT} \tag{6}$$

- Picos de potência do campo global (μV): quantidade de microestados de uma classe (picos de potência de campo global, GFPP) dividido pelo tempo total (em segundos) em que essa classe está presente no período analisado, exposto na equação (7);

$$AMP = \frac{GFPP}{CON} \tag{7}$$

- Correlação espacial média: métrica que analisa o ajuste espacial entre o mapa do microestado e as amostra de EEG atribuídas. Essa relação é calculada pela equação (8).

$$Corr_k = \frac{1}{N_k} \sum_{n}^{N_k} Corr\ (a_k, x_n) \tag{8}$$

onde a_k é o mapa protótipo para o *k*-ézimo microestado e N_k é a quantidade de amostras rotuladas.

2.7 *Backfit*

Após gerar o GFP a partir dos sinais de EEG e fazer a segmentação em microestados, é necessário realizar o *backfit* nos sinais dos pacientes com os protótipos gerados. Essa etapa consiste em mapear um sinal de EEG para que ele seja descrito como sequências dos microestados obtidos. Para que a representação do sinal seja a melhor possível, calcula-se o GMD (do inglês, *global map dissimilarity*) que representa a similaridade entre o sinal do EEG e a topografia do protótipo (POULSEN et al., 2018). Para duas amostras de EEG, x_i e x_j, obtém-se o valor do GMD a partir da equação (9).

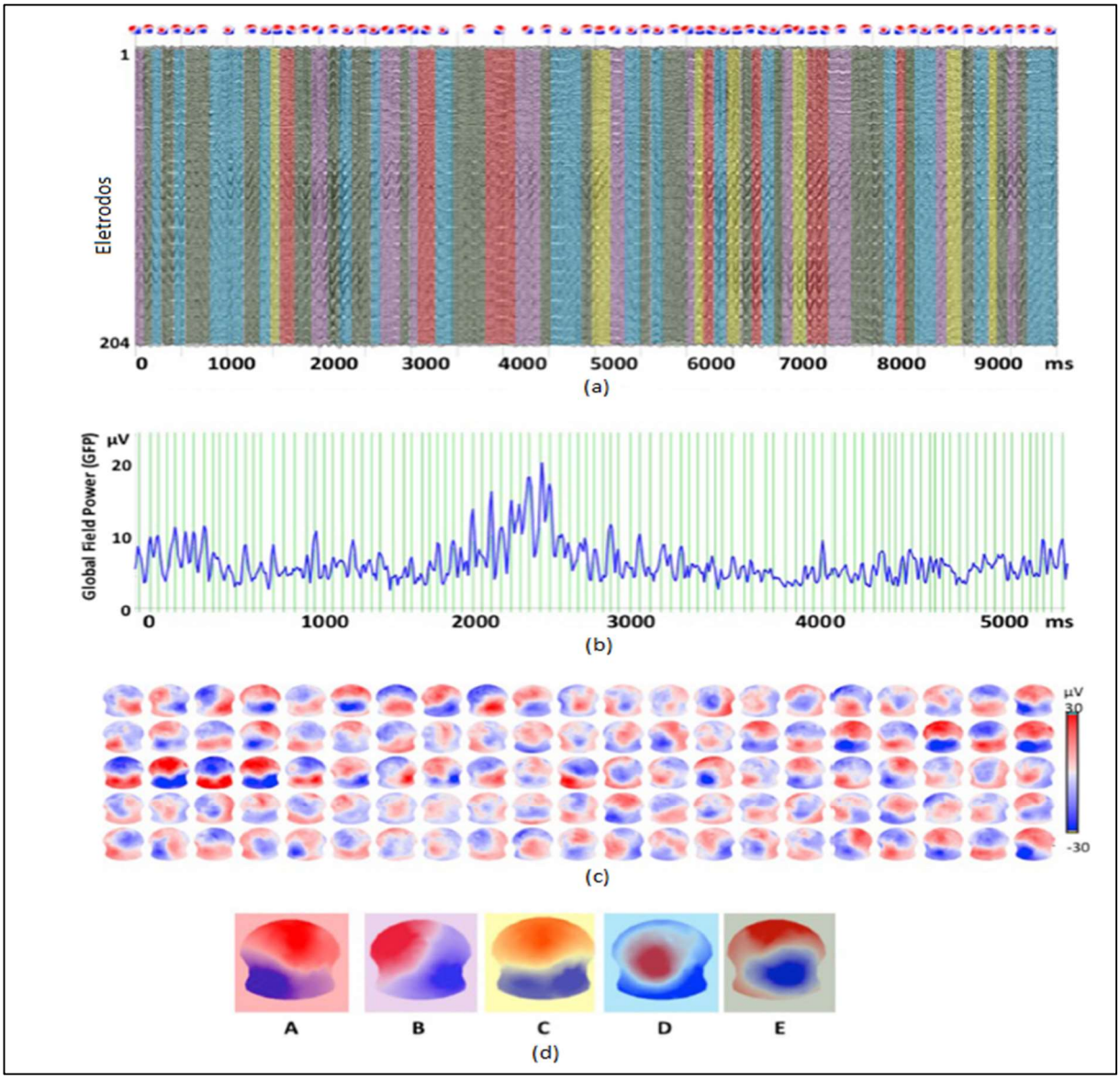
$$GMD = \frac{\left\| \frac{x_i}{GFP_i} - \frac{x_j}{GFP_j} \right\|}{\sqrt{C}} \tag{9}$$

onde *C* é a quantidade de eletrodos captando o sinal de EEG. O GMD é uma medida de distância que calcula quão similares são os mapas topográficos, sem se preocupar com a potência do sinal.

A Figura 10 ilustra como é o processo de obtenção dos microestados e como o sinal original de EEG é rotulado. Primeiramente, a partir dos sinais captados dos eletrodos (Figura 10(a)), calcula-se o sinal GFP (Figura 10(b)). Em seguida, são identificados os picos do sinal de GFP e, a partir deles, são determinadas as diversas topografias de microestados para cada pico

(Figura 10(c)). Depois, são estabelecidos quais os protótipos que melhor representam as várias topografias (Figura 10(d)). Por fim, estes protótipos são utilizados para rotular os sinais originais. As faixas de cores nos sinais de EEG na Figura 10(a) sinalizam quais dos microestados finais (protótipos) são utilizados para representar aquele trecho do sinal (MICHEL; KOENIG, 2018).

Figura 10 – Processo de segmentação de microestados: captação do sinal (a), cálculo do GFP (b), obtenção das topografias (c), determinação dos microestados (d)



Fonte: Michel e Koenig (2018).
Nota: Adaptado pelo autor.

3 METODOLOGIA

A metodologia adotada neste trabalho é caracterizada como sendo de abordagem quantitativa e natureza aplicada, pois tem como objetivo utilizar métodos de aprendizado de máquinas e análises estatísticas de dados para a obtenção de resultados de um problema específico: a previsão de crises convulsivas em pacientes epilépticos. Quanto aos objetivos, ela é definida como sendo do tipo de pesquisa explicativa, cuja finalidade é a exploração dos dados de sinais de eletroencefalogramas para encontrar padrões que permitam a previsão de futuras crises convulsivas. Em relação aos procedimentos, possui caráter experimental, onde variáveis decorrentes das análises estatísticas serão analisadas e testadas na busca de resultados que permitam a detecção do estado preditivo da convulsão epiléptica.

3.1 Descrição do Banco de Dados

O banco de dados utilizado é o CHB – MIT, que contém registros coletados no *Children's Hospital Boston* (SHOEB*,* 2009). Ele é um banco de dados público que consiste em gravações do EEG adquiridos de diferentes regiões do escalpo de 22 pacientes pediátricos com convulsões intratáveis. Os indivíduos foram monitorados por vários dias após a retirada da medicação anticonvulsivante e tiveram as suas crises convulsivas anotadas com horário de início e fim. O banco de dados não fornece o estado dos pacientes durante a obtenção dos sinais.

Todos os sinais de EEG são amostrados a 256 amostras por segundo e apresentam 23 canais bipolares, com gravações de 1, 2 ou 4 horas. Para essas gravações, o Sistema Internacional 10-20 foi utilizado para posições e nomenclatura dos eletrodos. Assim, os canais 1 a 23 são respectivamente: Fp1-F7, F7-T7, T7-P7, P7-O1, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F4, F4-C4, C4-P4, P4-O2, Fp2-F8, F8-T8, T8-P8, P8-O2, Fz-Cz, Cz-Pz, P7-T7, T7-FT9, FT9-FT10, FT10-T8 e T8-P8.

Conforme o estudo anterior de classificação dos estados interictais e preictais de Kill e outros (2019), para selecionar os estados interictais, são selecionados intervalos de 40 minutos, aproximadamente, 4 horas antes ou depois das crises epilépticas (estado ictal). Para o pré-ictal, são escolhidos intervalos de 40 min, até 1 segundo antes do ictal. Desse modo, apenas 20

pacientes possuem gravações que atendem a esses requisitos e são utilizadas na pesquisa (Quadro 2).

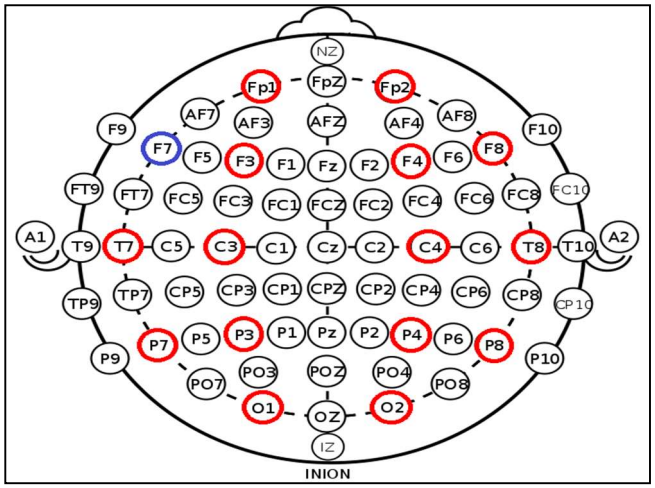
Quadro 2 – Pacientes do *Children's Hospital Boston* analisados neste trabalho

Paciente	Sexo	Idade (anos)
1	Masculino	11
2	Masculino	11
3	Feminino	14
4	Masculino	22
5	Feminino	7
6	Feminino	1.5
7	Feminino	14.5
8	Masculino	3.5
9	Feminino	10
10	Masculino	3
11	Feminino	12
12	Masculino	16
13	Feminino	9
14	Feminino	7
15	Feminino	12
16	Feminino	18
17	Feminino	19
18	Feminino	6
19	Feminino	3
20	Feminino	9

Fonte: Produzido pelo próprio autor.

Os dados são importados através do *software Matlab* e, para o estudo de microestados, são processados usando a caixa de ferramentas EEGLAB. Neste trabalho foram feitas adaptações no código de Poulsen e outros (2018) para que pudesse ser possível simular um método *on-line*. Os canais inicialmente bipolares são transformados em monopolares, calculando a diferença de potencial de cada par de eletrodos, usando como referência o canal F7 (circulado em azul na Figura 11) que foi escolhido aleatoriamente. Esse procedimento é realizado porque o EEGLAB requer o uso de canais monopolares com uma referência comum. Entretanto, a referência média é mais comumente utilizada em análise de microestados para gravações do EEG espontâneo, por isso, os sinais são referenciados, novamente, para a referência média com os 15 canais monopolares formados (circulados em vermelho na Figura 11) (POULSEN et al., 2018).

Figura 11 – Canais utilizados na análise de microestados



Fonte: Produzido pelo próprio autor.

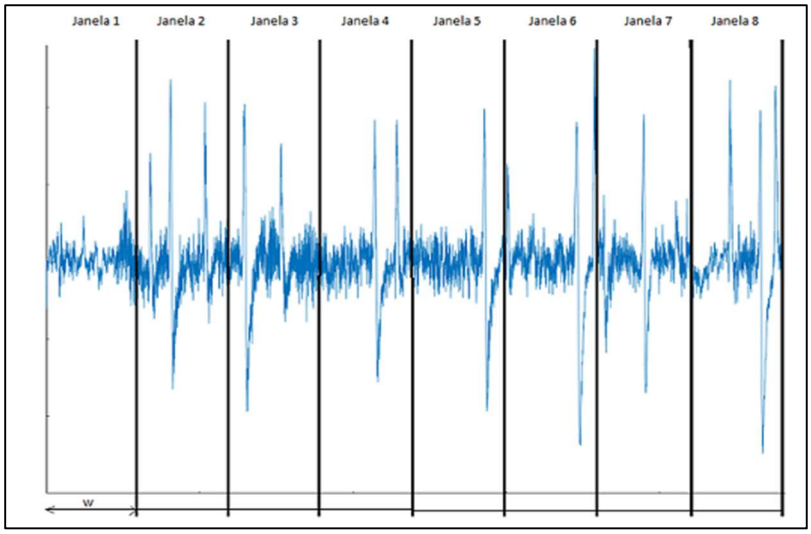
3.2 Pré-Processamento

O sinal do EEG é uma série temporal de uma sequência de valores reais, que são registrados ao longo do tempo. Inicialmente, os sinais são filtrados usando um filtro *notch* de 60 Hz para reduzir o ruído da rede de alimentação do equipamento. Para a análise de microestados, estudos anteriores mostraram que a segmentação do sinal do EEG com menos de 60 segundos de dados resultam em topografias estáveis do couro cabeludo (BARADITS; BITTER; CZOBOR, 2020).

A série temporal é então dividida em um conjunto de subsequências, denominados aqui de segmentos, usando uma função de janela deslizante. Uma série temporal é dividida em segmentos de tamanho fixo de comprimento w usando a função de janela deslizante, convertendo o sinal em segmentos, conforme a Figura 12. A análise é realizada em segmentos de w = 40 segundos de dados de EEG, uma vez que estudos anteriores mostraram que isso satisfaz os requisitos de qualidade para a análise de microestado (ANDREOU et al., 2014; BARADITS; BITTER; CZOBOR, 2020).

Desse modo, para cada paciente, os registros de 40 minutos interictais e de 40 minutos preictais foram divididos em 60 segmentos de 40 segundos cada, para os dados interictais e preictais. Como são 20 pacientes analisados, tem-se 1200 segmentos de 40 segundos interictais e 1200 segmentos de 40 segundos preictais, que são posteriormente processados simulando um método *on-line* com a caixa de ferramentas EEGLAB.

Figura 12 – Exemplo de uma série temporal dividida em 8 segmentos



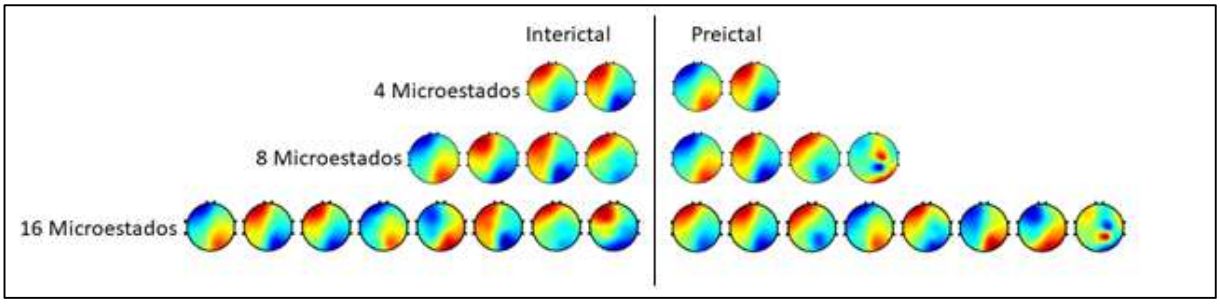
Fonte: Produzido pelo próprio autor.

3.3 Análise em Microestados

Para a análise de microestados foi seguido o procedimento padrão para segmentação de microestados apresentado em Poulsen et al. (2018). Para obtenção dos microestados, 40 segmentos de 40 segundos do estado interictal e 40 segmentos de 40 segundos do estado preictal são carregados ao EEGLAB. Em seguida, para cada conjunto (interictal e preictal) de dados, é calculado o GFP dos segmentos. Após isso, a *toolbox* retorna uma faixa de quantidades de microestados a se utilizar. A quantidade de microestados utilizada foi escolhida aleatoriamente para avaliar o impacto da quantidade de protótipos nos resultados obtidos pelos classificadores. Encontrar o número ideal de microestados é crucial para capturar os atributos informativos dos dados e evita ajustes excessivos/insuficientes (AHMADI et al., 2020).

Neste trabalho, o número ideal de microestados será avaliado de acordo com o critério de maior acurácia encontrada na classificação dos estados interictal e preictal. As análises são feitas com 4 microestados (2 interictais e 2 preictais), 8 microestados (4 interictais e 4 preictais) e 16 microestados (8 interictais e 8 preictais), conforme mostra a Figura 13. Para cada quantidade de protótipos, serão feitas classificações utilizando as características estatísticas de cada microestado, obtidas através do EEGLAB.

Figura 13 – Microestados criados para análise do número ideal de microestados



Fonte: Produzido pelo próprio autor.

Uma vez que os mapas de agrupamento global foram determinados, eles são ajustados para o sinal EEG de cada segmento de acordo com os valores de GFP(t) correspondentes para definir os microestados e assim, extrair diferentes características entre eles. Nessa etapa, os microestados são ajustados aos 1160 segmentos de 40 segundos interictal e aos 1160 segmentos de 40 segundos preictais, segmentos que não foram utilizados na etapa de obtenção dos microestados.

O procedimento de ajuste atribui rótulos de microestados a pontos de dados do EEG com base em qual mapa de cluster eles são mais topograficamente semelhantes com o uso da medida de dissimilaridade de mapa global (GMD). Tal procedimento permite representar a série temporal do EEG em termos de sequências de microestados e assim é possível extrair características estatísticas para cada microestado, como duração, ocorrência e cobertura (AHMADI et al., 2020).

3.4 Extração e Seleção de Características

Um total de 24 atributos (considerando 4 microestados), 48 atributos (utilizando 8 microestados) e 96 atributos (usando 16 microestados) são analisados neste trabalho. Esses atributos são calculados pelo EEGLAB para cada microestado, de acordo com as equações (2) a (7). A seleção de características é de considerável importância em aprendizado de máquina, especialmente para dados de alta dimensionalidade. Neste trabalho foi usado o algoritmo NCA, um método não paramétrico que é amplamente insensível ao aumento no número de atributos irrelevantes e tem um desempenho melhor do que outros métodos na maioria dos casos (WEI et al., 2012).

3.5 Modelos de Classificação

Os atributos que foram selecionados pelo NCA são então alimentados nos classificadores SVM e kNN para discriminação dos estados interictais e preictais. A validação cruzada *hold-out* (separando os dados em 80% para treino e 20% para teste) é utilizada para testar a generalização do modelo enquanto que a *k-fold* atua na validação e otimização dos parâmetros dos modelos. O procedimento de *hold-out* foi aplicado por 10 vezes enquanto que para o *k-fold* foi definido um valor de k = 5. O resultado final é a média dos resultados obtidos para cada métrica de desempenho: acurácia, sensibilidade e especificidade.

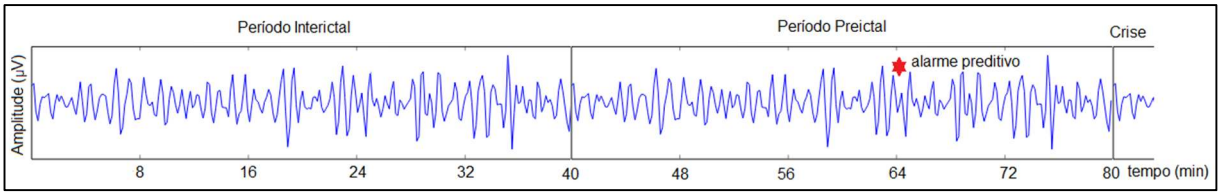
3.6 Método de Previsão da Crise Epiléptica

A previsão da crise é conduzida somente com os atributos selecionados pelo NCA e utilizando o classificador que apresentou melhor acurácia na etapa de classificação. Para determinar o tempo de previsão da antecedência de uma crise, o alarme preditivo, que informa que uma crise epiléptica está para acontecer, é acionado após a identificação do período preictal por 7 segmentos consecutivos de 40 segundos para aumentar a precisão do algoritmo e diminuir os falsos positivos. Na Figura 14 tem-se um sinal de 80 minutos, sendo os primeiros 40 minutos do período interictal e os 40 minutos seguintes do preictal, e uma representação do alarme preditivo. O uso de uma sequência de classificações do preictal aumenta a confiança nos resultados, pois se um classificador tem acurácia de 80%, por exemplo, a probabilidade de errar 7 vezes em seguida a classificação de um sinal interictal como preictal é menor do que 0,01% ($0{,}2^7$), enquanto que a probabilidade de classificar corretamente 7 vezes em seguida o sinal de preictal é de 20,97% ($0{,}8^7$). Embora a probabilidade seja pequena, se considerar os primeiros 20 minutos do sinal preictal, serão analisadas 30 janelas de 40 segundos e a probabilidade de identificar corretamente ao menos 7 vezes em seguida o sinal preictal em 30 janelas é de aproximadamente 84%.

O método de validação utilizado é o *leave-one-out*, em que o classificador é treinado com todos os segmentos de 19 pacientes e o teste é realizado com todos os segmentos do paciente retirado do conjunto de treinamento, simulando uma análise *on-line*. Sinais contínuos do EEG são selecionados e ordenados, contendo 40 minutos interictais seguidos de 40 minutos preictais,

com a finalização do preictal 1 segundo antes da crise. Esses períodos estão segmentados em janelas de 40 segundos.

Figura 14 – Sinal contínuo de EEG



Fonte: Produzido pelo próprio autor.

4 RESULTADOS

A classificação dos estados interictais e preictais foi realizada para 3 métodos de análise de microestados diferentes, a partir da obtenção de 4, 8 e 16 microestados. Para cada quantidade de microestado utilizada, o algoritmo foi executado 10 vezes, dividindo as amostras em 80% para treino (928 amostras) e 20% para teste (232 amostras). A seleção dos atributos ocorreu aplicando-se o NCA para fazer uma avaliação 5-*fold cross validation* sobre os dados de treino com o objetivo de determinar os melhores atributos. Em seguida, os mesmos 5-*fold* são utilizados para fazer a calibração dos hiperparâmetros de cada algoritmo, encontrando os melhores parâmetros para cada classificador. Nesta etapa, variou-se o valor de k do kNN de 1 a 15, utilizando apenas os números ímpares, e para os classificadores SVM linear e RBF foram usados valores do parâmetro de custo C de 0,25 a 2, em intervalos de 0,25. Em ambos os casos, foi utilizada a otimização da métrica acurácia, já que a base de dados é equilibrada. Nos Quadros 3 a 5 tem-se a média dos resultados das métricas de desempenho obtidos por cada classificador sobre o conjunto de teste depois das 10 execuções.

Quadro 3 – Resultados dos modelos utilizando atributos de 4 microestados

Classificador	Acurácia (%)	Especificidade (%)	Sensibilidade (%)
kNN	86,40	85,60	87,19
SVM linear	67,58	76,25	58,92
SVM RBF	**89,35**	**89,69**	**89,00**

Fonte: Produzido pelo próprio autor.

Quadro 4 – Resultados dos modelos utilizando atributos de 8 microestados

Classificador	Acurácia (%)	Especificidade (%)	Sensibilidade (%)
kNN	89,93	91,07	88,79
SVM linear	68,44	71,33	65,56
SVM RBF	**91,14**	**92,54**	**89,74**

Fonte: Produzido pelo próprio autor.

Quadro 5 – Resultados dos modelos utilizando atributos de 16 microestados

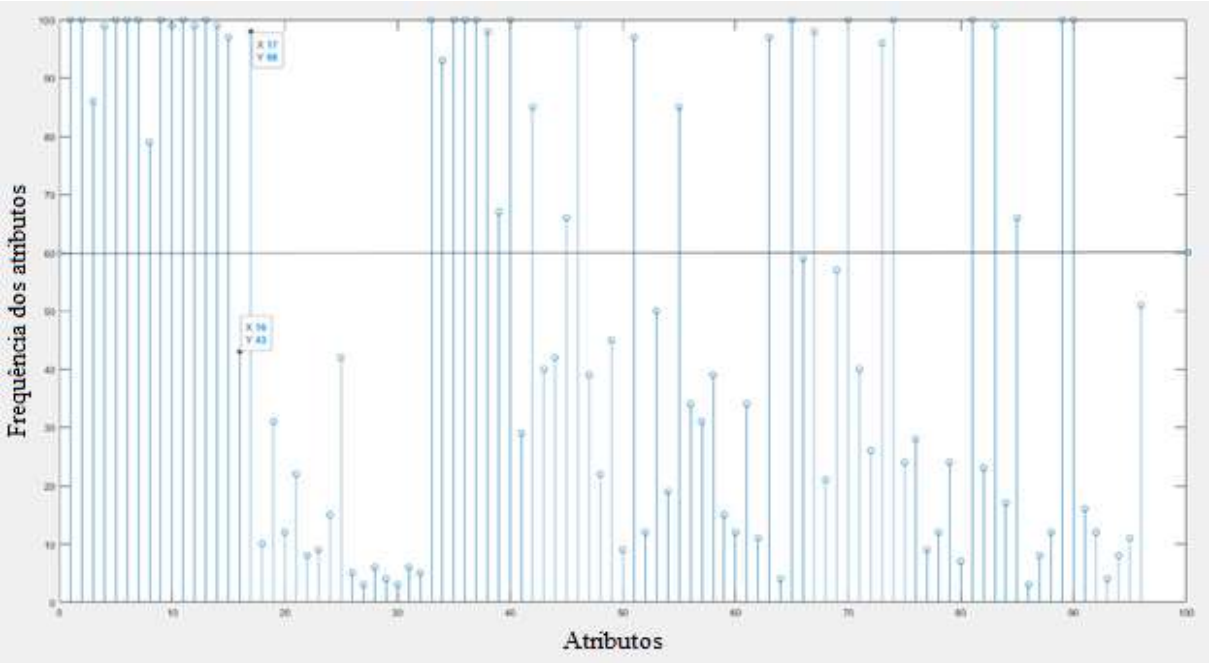
Classificador	Acurácia (%)	Especificidade (%)	Sensibilidade (%)
kNN	90,32	91,25	89,39
SVM linear	68,92	73,23	64,61
SVM RBF	**91,35**	**91,85**	**90,86**

Fonte: Produzido pelo próprio autor.

A partir dos resultados obtidos, observa-se que o SVM RBF se apresentou como melhor classificador neste estudo, alcançando aproximadamente 91,4% de acurácia utilizando 16 microestados.

Uma análise sobre os atributos que contribuem de forma mais significativa com a distinção dos estados interictal e preictal, e que se apresentaram mais recorrentes nos resultados do NCA, também foi feita. O algoritmo de seleção de características foi executado 100 vezes utilizando as informações de todos os 96 atributos, referentes à análise com 16 microestados. A observação foi feita com os atributos que apareceram mais de 60 vezes como saída do algoritmo. A Figura 15 mostra o gráfico da ocorrência dos atributos em cada iteração, sendo a linha horizontal um limiar para selecionar aqueles que apareceram mais de 60 vezes como saída do algoritmo e o Quadro 6 mostra quais são esses atributos.

Figura 15 – Representação gráfica da quantidade de vezes que cada atributo apareceu no resultado do NCA



Fonte: Produzido pelo próprio autor.

Quadro 6 – Representação dos atributos e seus respectivos microestados

Atributos	Microestados															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ocorrência	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
Duração	X															
Cobertura	X	X	X	X	X	X	X	X		X			X	X		
GEV			X				X								X	
GFP	X		X			X			X	X						
Correlação espacial média	X		X		X				X	X						

Fonte: Produzido pelo próprio autor.
Nota: Os atributos que estão marcados com um "X" estavam presentes no mínimo 60 vezes nos resultados do algoritmo NCA.

A partir dessa análise, entre os 96 atributos totais, apenas 40 demonstraram ter maior peso para uma boa classificação, como pode ser observado com mais detalhe no Quadro 2. Os atributos de duração média e do GEV mostraram serem medidas pouco eficientes na distinção dos estados interictal e preictal. Também se observa que os atributos associados aos primeiros microestados do interictal (1 e 3) e do preictal (9 e 10) são os mais frequentemente selecionados. Neste estudo não foi abordado o possível motivo dos atributos desses microestados possuírem maior relevância para a classificação. Desse modo, uma nova classificação foi feita utilizando os atributos mais frequentes das medidas de ocorrência, cobertura, GFP e correlação espacial média, a média dos 10 resultados é mostrada no Quadro 7.

Quadro 7 – Resultados dos modelos utilizando 16 microestados e os atributos selecionados pelo NCA

Classificador	Acurácia (%)	Especificidade (%)	Sensibilidade (%)
kNN	91,85	92,24	91,46
SVM linear	67,11	72,15	62,07
SVM RBF	**92,37**	**92,58**	**92,15**

Fonte: Produzido pelo próprio autor.

Como pode ser observado, os resultados do SVM RBF e kNN apresentados no Quadro 7 melhoraram ligeiramente em relação aos obtidos no Quadro 6, indicando a relevância dos atributos selecionados.

A análise subsequente foi realizada com os 20 pacientes para verificar o tempo de antecedência para prever uma crise epiléptica, utilizando os 40 atributos mais recorrentes, cujos resultados estão mostrados no Quadro 8. O menor tempo de previsão foi de 13,33 minutos antes da crise, enquanto que o maior, desconsiderando os alarmes falsos, foi de 18,67 minutos. Para os pacientes 6, 7, 9 e 17 houve a presença de falsos negativos, uma vez que quaisquer sinais com mais de 40 minutos antes da crise são sinais interictais. Na Figura 16, pode-se observar as janelas classificadas pelo SVM RBF para os pacientes 1 e 2. Após detectar 7 segmentos seguidos como pertencentes à classe preictal, dispara-se o alarme preditivo.

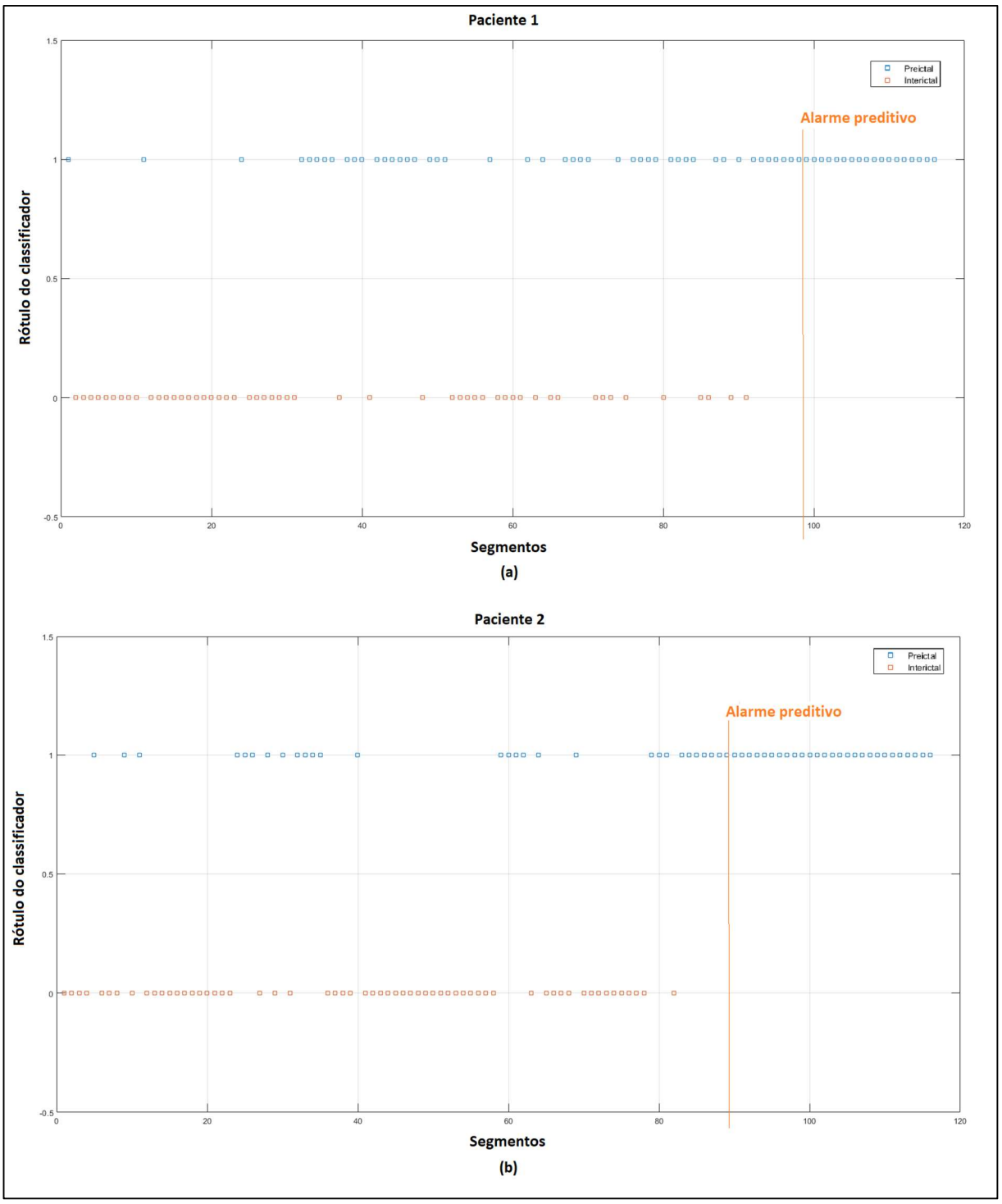
Quadro 8 – Resultados das predições para os 20 pacientes

Paciente	Tempo de predição antes da crise (em minutos)
1	13,33
2	18,67
3	36,64 / 27,97 / 17,97
4	36,64 / 27,31 / 21,31 / 9,97
5	36,64 / 23,97 / 17,97
6	59,97 / 39,31 / 3,97
7	71,31 / 31,31 / 25,31
8	27,31
9	63,97 / 36,64 / 5,97
10	36,64 / 19,31
11	30,64 / 19,31
12	36,64 / 23,31 / 12,64 / 7,31
13	33,97 / 8,64
14	36,64 / 9,31
15	36,64 / 17,97
16	36,64 / 5,31
17	59,31 / 51,31 / 35,31 / 19,97 / 5,97
18	31,97 / 8,64
19	38,64 / 25,97 / 8,64
20	15,97

Fonte: Produzido pelo próprio autor.

Apesar da literatura fornecer valores maiores de sensibilidade (THARA; PREMASUDHA; XIONG, 2019), o método aqui proposto apresenta uma identificação prévia do período preictal com alta precisão da breve ocorrência de uma convulsão, e com um tempo de previsão suficiente para que o mesmo se coloque em segurança ou tome os medicamentos necessários.

Figura 16 – Sinal dos pacientes 1 (a) e 2 (b) segmentados em janelas de 40 segundos e classificados pelo SVM RBF



Fonte: Produzido pelo próprio autor.

Comparando o melhor resultado deste trabalho, com os resultados de outro estudo que utiliza a análise de microestados para epilepsia (RAJ et al., 2018), pode-se observar que este projeto obteve valores superiores para as métricas de desempenho (Quadro 9). Vale ressaltar, no

entanto, que as bases de dados utilizadas foram diferentes, podendo ser um motivo que impactou nos valores de acurácia, sensibilidade e especificidade.

Quadro 9 – Comparação de resultados obtidos por Raj e outros (2018) e o melhor classificador deste trabalho

Algoritmo	Acurácia (%)	Especificidade (%)	Sensibilidade (%)
Análise discriminante linear de Fisher (RAJ et al.; 2018)	73,80	66,70	81,00
Regressão logística (RAJ et al., 2018)	66,70	71,40	61,90
SVM RBF	**92,37**	**92,58**	**92,15**

Fonte: Produzido pelo próprio autor.

5 CONCLUSÕES

Neste trabalho foi apresentado um estudo de um projeto de aprendizado de máquinas voltado para a área da saúde. Nele, são explicitados diversas etapas de projetos de AM, sendo eles: a obtenção dos dados, o pré-processamento, a extração de características, a construção de modelos e suas otimizações até, por fim, chegar em um resultado conclusivo.

Durante o desenvolvimento, foram feitas as devidas modificações no algoritmo de Poulsen e outros (2018) para que uma análise *on-line* de sinais de EEG de pacientes epilépticos pudesse ser realizada. Esse feito tem capacidade de auxiliar no surgimento de novas pesquisas focadas na análise de microestados para a classificação de outras doenças neurológicas, visto que a detecção *on-line* é, em muitos dos casos, bem significativa para uma possível aplicação.

A *toolbox* EEGLAB, desenvolvida por Poulsen e outros (2018) para o *software Matlab*, mostrou-se útil ao longo do projeto. A capacidade de geração do sinal de GFP, junto com a segmentação em microestados e *backfit* dos sinais de EEG são apenas algumas das diversas ferramentas presentes no algoritmo. A extração dos dados estatísticos de cada microestado, feito pelo próprio EEGLAB, foi fundamental para a criação dos modelos, pois, usando-os como características distintas, foi possível treinar, validar e testar os classificadores. Além disso, foram usadas técnicas para a obtenção do melhor conjunto de atributos, visando a otimização dos resultados obtidos.

Os resultados finais, apresentados nos Quadros 3 a 5 e Quadro 7, foram considerados satisfatórios. Analisou-se três tipos de classificadores, amplamente utilizados na literatura, neste projeto: o kNN, o SVM linear e o SVM RBF. O algoritmo SVM linear foi o que apresentou o resultado inferior, obtendo uma acurácia máxima de 68,92% (8 microestados), especificidade máxima de 76,25% (4 microestados) e sensibilidade máxima de 65,56% (4 microestados), valores baixos considerando-se se tratar de uma aplicação voltada para a área da saúde. Os algoritmos kNN e SVM RBF, no entanto, apresentaram uma acurácia máxima de 91,85% e 92,37, especificidade máxima de 92,24 % e 92,58% e sensibilidade máxima de 91,46% e 92,15%, respectivamente (resultados obtidos utilizando 16 microestados e os atributos selecionados pelo NCA).

Cita-se também que a quantidade de microestados utilizado na análise afeta ligeiramente os resultados obtidos. Contudo, foi notada uma tendência na melhora dos resultados, conforme aumentou-se o número de atributos para os classificadores kNN e SVM RBF, porém o mesmo não pode ser afirmado para o SVM linear. Uma análise feita sobre os atributos obtidos de 16 microestados indica que nem todas as medidas estatísticas usadas como atributos são úteis.

Adicionalmente, a previsão da crise para os 20 pacientes foi realizada com o tempo de previsão acima de 12 minutos antes do acontecimento da crise. O que mostra uma eficácia do método em facilitar a vida dos pacientes com um alarme que fornece tempo suficiente para se colocarem em uma situação segura ou se prevenirem por meio de medicamentos. Análises futuras poderão ser realizadas com maior número de pacientes.

Uma sugestão para futuros trabalhos para a análise de microestados é a obtenção de novas características para se utilizar na construção dos modelos. Ao se realizar o *backfit*, mapeia-se o sinal de EEG em vetores rotulados de acordo com o microestado atribuído. Desse vetor, é possível obter-se novos atributos que podem auxiliar na melhora dos resultados das classificações. Além disso, pode-se utilizar métodos, aplicados em projetos de aprendizado de máquinas, chamados de Engenharia de Atributos, que consiste em manipular atributos já obtidos para criar novas características para alimentar o modelo.
Outra sugestão seria na exploração de outros classificadores não abordados durante este estudo. Conforme mostrado, diferentes classificadores obtiveram diferentes resultados utilizando os mesmo dados para treino e teste. Pode-se realizar um estudo que explora os diferentes modelos presentes na literatura na busca por um que apresente melhores resultados que os obtidos neste projeto.

A última sugestão seria o desenvolvimento de uma métrica para avaliar a semelhança entre microestados. Durante a execução deste trabalho, na etapa de obtenção de microestados, é possível perceber que algumas topografias são parecidas. Um futuro estudo pode analisar se esses microestados equivalentes podem ser representados por apenas uma única topografia, reduzindo a quantidade de protótipos utilizada.

REFERÊNCIAS BIBLIOGRÁFICAS

ACHARYA, U. R.; HAGIWARA, Y.; ADELI, H. Automated seizure prediction. **Epilepsy & Behavior**, v. 88, p. 251-261, nov. 2018.

AHMADI, N.; PEI Y.; CARRETTE, E.; ALDENKAMP A. P.; PECHENIZKIY, M. EEG-based classification of epilepsy and PNES: EEG microstate and functional brain network features. **Brain Informatics** v. 7, n. 6, jun. 2020.

ANDREOU, C.; FABER, P. L.; LEICHT, G.; SCHOETTLE, D.; POLOMAC, N.; HANGANU-OPATZ, I. L.; LEHMANN, D.; MULERT, C. Resting-state connectivity in the prodromal phase of schizophrenia: insights from eeg microstates. **Schizophrenia Research**, v. 152, n. 2-3, p. 513-520, fev. 2014.

ASSI, E. B.; NGUYEN, D. K.; RIHANA, S.; SAWAN, M. Towards accurate prediction of epileptic seizures: A review. **Biomedical Signal Processing and Control**, v. 34, p. 144-157, abr. 2017.

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Machine Language**, v. 20, n. 3, p. 273-297, set. 1995.

DEVINSKY, O.; VEZZANI, A.; O'BRIEN, T. J.; JETTE, N.; SCHEFFER, I. E.; CURTIS, M; PERUCCA, P. Epilepsy. **Nature Reviews Disease Primers**, v. 4, n. 1, artigo 18024, jun. 2018.

FU, Y.; CHEN, J.; XIONG, X. Calculation and Analysis of Microstate Related to Variation in Executed and Imagined Movement of Force of Hand Clenching. **Computational Intelligence and Neuroscience**, v. 2018, artigo 9270658, ago. 2018.

GODFREY, K. Simple Linear Regression in Medical Research. **New England Journal of Medicine**, v. 313, p. 1629-1636, dez. 1985.

GOLDBERGER, J.; ROWEIS, S.; HINTON, G.; SALAKHUTDINOV, R. Neighborhood Components Analysis. **Advances in Neural Information Processing Systems**, v. 17, p. 513-520, dez. 2004.

GUO, G.; WANG, H.; BELL, D.; BI Y.; GREER K. KNN Model-Based Approach in Classification. **Lecture Notes in Computer Science**, v. 2888, p. 986-996, jan. 2003.

HANOUSEK, J. **Neighborhood components analysis and machine learning**. 2018. Tese (Doutorado em Matemática) – Department of Probability and Mathematical Statistics, Charles University, Praga, 2018.

KHANNA, A.; PASCUAL-LEONE, A.; MICHEL, C. M.; FARZAN, F. Microstates in resting-state EEG: Current status and future directions. **Neuroscience & Biobehavioral Reviews**, v. 49, p. 105-113, fev. 2015.

KILL, J. B.; CIARELLI, P. M.; CÔCO, K. F.; SOUZA, M. L. Wavelet Analysis Applied on EEG Signals for Identification of Preictal States in Epileptic Patients/Análise wavelet aplicada em sinais de EEG para identificação de estados pré-letais em pacientes epilépticos. **Brazilian Applied Science Review**, v. 4, n. 3, p.1730-1747, out. 2019.

MICHEL, C. M.; KOENIG, T. EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: A review. **NeuroImage**, v. 180, p. 577-593, out. 2018.

MISHRA, A.; ENGLITZ, B.; COHEN, M. X. EEG Microstates as a Continuous Phenomenon. **NeuroImage**, v. 208, artigo 116454, mar. 2020.

MORENO-TORRES, J. G.; SAÉZ, J. A.; HERRERA, F. Study on the Impact of Partition-Induced Dataset Shift on *k*-fold Cross-Validation. **IEEE Transactions on Neural Networks and Learning Systems**, v. 23, n. 8, p. 1304-1312, ago. 2012.

MORMANN, F.; ANDRZEJAK, R. G.; ELGER, C. E.; LEHNERTZ, K. Seizure Prediction: The Long and Winding Road. **Brain**, v. 130, n. 2, p. 314-333, fev. 2007.

NJOGHOLO, B. M. **Investigating Keystroke Dynamics as a Two-Factor Biometric Security**. 2018. Dissertação (Mestrado em Ciência em Segurança de Sistemas de Informação) – Faculty of Information Technology (FIT), Strathmore Univesity, Nairobi, Kenya, 2018.

NISHIDA, K.; MORISHIMA, Y.; YOSHIMURA, M.; ISOTANI, T.; IRISAWA, S.; JANN, K.; DIERKS, T.; STRIK, W.; KINOSHITA, T.; KOENIG, T. EEG microstates associated with salience and frontoparietal networks in frontotemporal dementia, schizophrenia and Alzheimer's disease. **Clinical Neurophysiology**, v. 124, n. 6, p. 1106-1114, jun. 2013.

NUNEZ, P.; SRINIVASAN, R. Electroencephalogram. **Scholarpedia**, vol. 2, n. 2, p. 1348 2007. Disponível em: http://www.scholarpedia.org/article/Electroencephalogram. Acesso em: 29 maio 2020.

OLIVEIRA JUNIOR, G. M. **Máquinas de Vetores de Suporte: estudo e análise de parâmetros para otimização de resultado**. 2010. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Centro de Informática, Universidade Federal de Pernambuco, Recife, 2010. Disponível em: https://www.cin.ufpe.br/~tg/2010-2/gmoj.pdf. Acesso em: 15 ago. 2020.

PIORECKA, V.; PIORECKY, M.; STROBL, J.; NEZBEDOVA, M.; SCHAABOVA, H.; KRAJČA, V. EEG microstates analysis in patients with epilepsy. **Lekar a Technika**, v. 48, p. 96-102, jan. 2018.

POULSEN, A. T.; PEDRONI, A.; LANGER, N.; HANSEN, L. K. Microstate EEGLAB toolbox: An introductory guide. **bioRxiv**, p. 1-30, mar. 2018. Disponível em: https://www.biorxiv.org/content/10.1101/289850v1. Acesso em: 01 maio 2020.

RAJ, K.; RAJAGOPALAN, S. S.; BHARDWAJ, S.; PANDA, R.; REDDAM, V. R.; CHAITANYA, G.; RAGHAVENDRA, K.; MUNDLAMURI, R. C.; THENNARASU, K.; MAJUMDAR, K. K.; SATISHCHANDRA, P.; SINHA, S.; BHARATH, R. D. Machine

Learning detects EEG microstate alterations in patients living with temporal lobe epilepsy. **Seizure: European Journal of Epilepsy**, v. 61, p. 8-13, out. 2018.
SHOEB, A. H. **Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment**. 2009. Tese (Doutorado em Filosofia da Engenharia Elétrica e Médica) – Harvart MIT Division of Health Sciences and Technology, Massachussets Institute of Technology, Massachussets, Cambridge, 2009. Disponível em: https://physionet.org/content/chbmit/1.0.0/. Acesso em: 10 maio 2020.

THARA, D. K.; PREMASUDHA, B. G.; XIONG, F. Epileptic seizure detection and prediction using stacked bidirectional long short term memory. **Pattern Recognition Letters**, v. 128, p. 529-535, 2019.

WONG, T. T. Performance evaluation of classification algorithms by *k*-fold and leave-one-out cross validation. **Pattern Recognition**, v. 48, p. 2839-2846, set. 2015.

WORLD HEALTH ORGANIZATION. **Epilepsy**. 2019. Disponível em: https://www.who.int/news-room/fact-sheets/detail/epilepsy. Acesso em: 28 maio 2020.